基于词频统计的蛋白质交互关系识别

蔡松成,牛 耘
(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘 要: 目前,基于远监督的蛋白质交互关系抽取方法通过将知识库中的实体对与文本中的实体进行匹配来产生大规模的训练数据,有效地解决了标注数据不足的问题。在基于最大期望算法的蛋白质交互识别的基础上,提出了一种基于词频统计的蛋白质交互关系识别。该方法对每一个蛋白质对签名档进行处理,取出两个目标蛋白质中间的单词;然后对其进行词性标注,只保留名词和动词,同时进行词干提取;最终得到每个蛋白质对签名档下的词频统计。利用得到的词频信息设定阈值来获取签名档的高频词,改进最大期望算法的初始化过程。实验结果表明,通过加入高频词信息的干预来进一步获取句子的类别作为初始值较原始的基于最大期望算法的模型,取得了更高且均衡的精确度和召回率,对目前基于远监督的蛋白质交互关系识别方法进行了明显的改进。

关键词: 远监督;蛋白质交互;最大期望算法;词频统计
中图分类号: TP391 **文献标识码:** A **文章编号:** 1673-629X(2019)02-0065-04
doi: 10.3969/j.issn.1673-629X.2019.02.013

Protein-protein Interaction Identification Based on Word Frequency Count

CAI Song-cheng, NIU Yun
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Current protein-protein interaction (PPI) extraction approach based on distant supervision gathers large scales of training data by aligning entity pairs in knowledge base with entities in text, which solves the problem of lack of annotation data effectively. In this paper, based on the protein interaction recognition using the expectation maximization algorithm, we propose a novel method of word frequency count, which processes the signature of each protein pair and obtains the unigram words between two target proteins. Then, the data which is obtained by the first step should be processed with POS tagging and stem extraction, only the nouns and verbs saved. Finally, we can obtain the word frequency statistics for signatures of protein pairs. High frequency words are produced by setting the threshold for the word frequency statistics, which can be used to improve the initialization step of the expectation maximization algorithm. The experiment shows that the high and well balanced precision and recall are achieved by further integrating the high-frequency word information to obtain the sentence category as the initial model based on the maximum expectation algorithm, which shows significant improvement in comparison to current PPI based on distant supervision.

Key words: distant supervision; protein-protein interaction; expectation maximization algorithm; word frequency count

0 引 言

通过相互作用,细胞中的蛋白质完成细胞中的大部分过程,比如细胞内通讯。因而,蛋白质交互信息(protein-protein interaction, PPI)成为了关键信息,用以解决大量医学难题。目前,生物学家通过人工阅读的方式识别医学文献中的 PPI,并按照统一的格式将这些重要的信息录入数据库,如 HPRD[1]、IntAct[2]、MINT[3]和 BIND[4]等。然而以上数据库中的 PPI 信息并不全面,而且生物医学的快速发展导致每年相关科学文献的增长数量达上千万,每天也在产生新的蛋白质之间的关系。因此要从医学文献中收集 PPI 信息,仅靠手工方式难以满足现实的需求。

此背景下,有监督的机器学习方法被大量地运用到研究 PPI 关系识别中,并取得了巨大进展。由于有

收稿日期: 2018-03-29 **修回日期:** 2018-08-02 **网络出版时间:** 2018-11-15
基金项目: 国家自然科学基金(61202132)
作者简介: 蔡松成(1994-),男,硕士研究生,研究方向为自然语言处理;牛 耘,副教授,CCF 会员(E200035388M),研究方向为自然语言处理。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181115.1051.088.html

监督的机器学习方法需要大量人工标注的数据,代价高昂,所以很多研究者将远监督[5]的思想应用到 PPI 识别上,以解决训练数据不足的问题。但是由于远监督思想的缺陷,引入了大量的噪音,影响现阶段 PPI 识别的精度。因此,有研究人员提出了一种基于最大期望算法的蛋白质交互识别方法,构建多实例多标记学习模型,有效消除了签名档中噪音对交互关系识别的影响。文中在该方法的基础上,对每一个蛋白质对的签名档进行词频统计,得到相应的高频词。利用这些信息对最大期望算法的初始化过程进行改进,继而进行 PPI 识别。

1 相关工作

随着机器学习的流行,研究者们越来越多地采用基于机器学习的方法进行 PPI 识别。基于机器学习的方法主要包括两大类: 基于核函数的方法和基于特征的方法。基于核函数的方法首先对句子结构进行深入研究,通过设计核函数衡量不同蛋白质对间的相似度,然后使用支持核函数的分类器进行 PPI 关系识别。例如,Bunescu R C 等[6]提出了最短依赖路径核; 文献[7-9]使用基于多核(特征的核,树核及图核融合)的学习方法抽取 PPI 信息。基于特征的方法试图从标注有交互关系的句子中抽取重要特征,包括词汇特征、语法特征和语义特征,建立模型来判断蛋白质之间的交互关系[10-13]。但有监督的方法需要大量有标注的数据,且研究对象是单个句子,因此只能依赖一个句子中的线索,对于复杂的句子描述很难判断。

2 基于最大期望算法的 PPI 识别

文中主要从以下几个方面对基于最大期望算法的多实例多标记学习框架[14-15]加以改进。

2.1 基于高频词的签名档句子筛选

以远监督为基础的关系抽取方法利用已有的知识库和文本集,通过启发式的匹配来提供训练数据。这种方法可以产生大量带标注的训练数据,很好地解决了人工标注数据代价昂贵的问题,节省了人力物力。但是这种匹配是基于假设条件: 如果文本集中的语句包含了知识库的实体对,那么这条语句就表达了实体对在知识库中所对应的关系。显然,该假设过于理想化,会产生大量的噪音数据。

作为关系抽取在生物信息学方面的应用,基于远监督的蛋白质交互关系抽取同样面临训练数据存在噪音的问题。首先,利用现有的 HPRD 数据库查询获取有交互关系的蛋白质对,然后从 PubMed 数据库中自动获取包含蛋白质对的句子形成签名档。在得到的签名档中,部分句子并未表明目标蛋白质对之间的交互关系。例如,(ahr,rela)是一对有交互关系的蛋白质对,下面是其签名档中的两个句子:

1: We demonstrated that #ahr# associates with #rela # in the cytosol and nucleus of human lung cells.

2: Thus, the #rela# and #ahr# proteins functionally cooperate to bind to NF-kappaB elements and induce c-myc gene expression.

在这两个句子中,第一个句子确实表达了目标蛋白质对(ahr,rela)之间的交互关系,而第二个句子只是包含了目标蛋白质对,但并未表达两者之间的交互关系,所以这个句子成为了该蛋白质对的噪音。训练数据中的这部分噪音会影响最终模型的性能。因此,文中提出了基于高频词的去噪模型来对训练数据中的噪音进行处理。

2.1.1 提取签名档的词频统计信息

通过对蛋白质对签名档数据的观察发现,对于这些描述蛋白质交互作用的句子在单词级别上是存在相似性的。句子中经常出现 bind、interact、activate、association、ligand、inhabit、induce 等表示蛋白质交互作用的单词。鉴于生物医学文本对于蛋白质交互关系表达的这种规律,试图对每一个蛋白质对的签名档进行词频统计,同时设定阈值,得到对应的高频词集。根据签名档的高频词集,对签名档中的句子进行去噪处理。

提取签名档的词频统计信息主要包括以下步骤:

(1) 针对签名档数据,利用 NLTK 自然语言处理工具包进行词性标注;

(2) 两个目标蛋白质之间的单词对描述交互关系更为重要,所以只对这部分单词进行词频统计;

(3) 进行单词预处理,删除长度小于 2 的以及为纯数字的单词,同时将单词中的大写字母转化为小写字母;

(4) 进行词频统计时,只考虑名词和动词这两种具有实际意义的单词,这两种词性的单词对蛋白质交互作用的描述起到了非常重要的作用,同时剔除掉名词中的专有名词;

(5) 考虑到选择出来的名词中可能包含除目标蛋白质以外的其他蛋白质,因此采用 abner 蛋白质命名实体工具,识别出其他蛋白质的名称,将其去除掉;

(6) 利用 NLTK 工具包进行英文词干提取。

通过以上 6 个步骤,最终得到每个蛋白质对签名档下的词频统计信息。

2.1.2 基于高频词的去噪模型

本节根据蛋白质对的高频词集对签名档中的句子进行去噪处理。

算法 1: 利用高频词集确定句子类别。

输入:

词频阈值 TC

词频统计映射表 WC

高频词(high frequency words) 集合 HFW = { } //高频词和频率的映射

句子特征词集合 $F=\{f_1, f_2, \cdots, f_n\}$

输出: 句子标签集 $label=\{r_1, r_2, \cdots, r_m\}$

1: if m<=C then
2: HFW = WC
3: else then
4: for count in WC. keys( ) do
5: if count>TC then
6: HFW = HFW ∪ { count: WC [count] }
7: flag = false //用以判断句子类别的辅助标记
8: for $f_k \in F_i$ do
9: if f_k in HFW. keys( ) then
10: r_i = '1'
11: flag = true
12: if flag = = flase then
13: r_i = '0'
14: return label

算法 1 描述了如何利用高频词集合确定句子的类别,其中句子特征词集合就是利用 2.2.1 节提取词频统计信息的 6 个步骤后,得到每个句子两个目标蛋白质中间的若干一元词。

算法 1 主要包含两个步骤:

(1) 如果签名档的大小小于阈值 C ,保留词频统计信息作为高频词集; 大于阈值 C ,则对于词频统计映射表中频率大于词频阈值 TC 的单词才被保存到高频词集中,对应算法 1 的 1 ~ 6 行,文中设置 TC 为 1。

(2) 根据得到的高频词集合来确定句子标签,如果句子的特征词集合中包含高频词,就认为其是有交互关系的,否则认为这个句子不表达交互关系,对应算法的 7 ~ 13 行。

接下来通过获取到的句子标签集合对句子进行去噪处理,算法 2 描述了基于高频词的句子筛选过程。

算法 2: 基于高频词的句子去噪处理。

输入:

有交互蛋白质对集 P

无交互蛋白质对集 N

句子标签集 $label=\{r_1, r_2, \cdots, r_m\}$

签名档中的句子集合 $S=\{s_1, s_2, \cdots, s_m\}$

输出: 签名档中的句子集合

1: if $<e_1, e_2> \in P$ then
2: for $r_i \in$ label do
3: if r_i = = '0' then
4: count_1+ = 1
5: if count_1 = = m then
6: randomly save somesentences, others remove from S
7: else then
8: for $r_i \in$ label do
9: if r_i = = '0' then
10: removes$_i$ from S
11: else then
12: for $r_i \in$ label do
13: if r_i = = '1' then
14: count_2+ = 1
15: if count_2 = = m then
16: randomly save some sentences and set the labe1 as '0'
17: else then
18: for $r_i \in$ label do
19: if r_i = = '1' then
20: remove s_i from S
21: return S

算法 2 对于有交互的蛋白质对,根据高频词集判断为无交互的句子被认定为噪音数据,要从签名档中去除,对应算法的 7 ~ 10 行。同时,可能一个蛋白质对下的所有句子标记全是无交互的,为了避免把该蛋白质对直接从训练数据中过滤掉,随机保留部分句子作为该蛋白质对的签名档,对应算法的 1 ~ 6 行。对于无交互蛋白质对签名档中的句子,根据先验知识,它们应该都不表达交互关系,但是通过高频词集,同样会把无交互蛋白质对中的部分句子认定为是有交互关系的,这部分句子成为无交互蛋白质对中的噪音,需要去除,对应算法的 17 ~ 20 行。同样,当整个签名档中的句子均判断为有交互时,随机保留部分句子。和有交互蛋白质对不同的是,对于这部分句子,根据先验知识,将其标记改为无交互的,对应算法的 12 ~ 16 行。

2.2　最大期望算法的模型训练

由于蛋白质对签名档中的句子标记是未知的,因此利用最大期望算法在模型存在未知变量的情况下,对模型参数进行极大似然估计。

2.3　最大期望算法的具体实现

在利用最大期望算法实现多实例多标记学习模型的过程中,为了保证模型的学习效果,主要针对最大期望算法的初始化步骤进行了处理。

初始化: 由于最大期望算法并不是全局最优算法,无法保证找到全局最优解,因此初始值的选择非常重要。在该模型中,初始值为签名档中句子的类别分布 z_i ,算法起始于 M 步,在 M 步利用初始句子类别训练模型中的分类器,然后在 E 步骤中对初始句子类别重新判断,重复迭代。文中主要设置以下两种初始值:

方案 C_1: 以远监督匹配得到的句子类别作为多实例多标记模型的初始值。既有交互蛋白质对签名档中的句子均为有交互,无交互签名档中的句子均视为无交互。

方案 C_2: 利用基于高频词集合去噪处理后得到的

句子及其句子类别作为多实例多标记模型的初始值。

文中采用这两种不同的初始值设置方式试图求得模型的最优解。

3 实 验

3.1 实验数据

实验中有交互关系的蛋白质对是直接从 HPRD 数据库中查询获取，并且只保留被 PubMed 数据库中一篇以上摘要包含的那些蛋白质对。而对于无交互关系的蛋白质对，将 HPRD 中的蛋白质随机组合成蛋白质对，去除已被 HPRD 数据库包含的蛋白质对以及未被 PubMed 数据库记载的蛋白质对。以一对蛋白质为查询参数，从文献中检索出描述这两个蛋白质的所有句子，作为该蛋白质对的签名档。最终总共得到有交互关系和无交互关系的蛋白质对分别为 576 对和 578 对，合计 1 154 对。

3.2 实验设置

实验采用的结果性能评价指标是当前 PPI 抽取系统主要使用的三个指标：精确度（Precision = TP/(TP+FP)）、召回率（Recall = TP/(TP+FN)）和 F 值（F-Score = $2P \times R/(P+R)$）。为了避免简单应用模型而产生过拟合问题，利用五折交叉验证来评估模型的性能。

3.3 实验结果及分析

为了比较最大期望算法的两种初始值设置方式的实验效果，取最大期望算法迭代的前 6 次（迭代 6 次以后实验结果基本趋向局部最优解），对有交互蛋白质对的识别结果如表 1、表 2 所示。

表 1 方案 C_1 的识别结果 %

迭代次数	精确度	召回率	F 值
$T=1$	70.8	67.0	69.0
$T=2$	67.4	70.0	68.4
$T=3$	69.8	72.0	70.8
$T=4$	68.8	74.0	71.2
$T=5$	66.8	75.0	70.6
$T=6$	65.0	75.4	69.8

表 2 方案 C_2 的识别结果 %

迭代次数	精确度	召回率	F 值
$T=1$	68.4	79.6	73.4
$T=2$	71.4	72.0	71.6
$T=3$	72.2	72.8	72.6
$T=4$	73.2	67.8	70.4
$T=5$	72.8	68.6	70.6
$T=6$	73.6	67.4	70.6

通过对表 1、表 2 的观察可以发现，在前三次迭代，实验方案 C_2 在精确度、召回率和 F 值上都有明显的提高。随着迭代次数的增加，C_2 获得了较高的查准率，但是查全率相对较低。说明在基于高频词的去噪模型中，由于存在一部分被判断为噪音数据的句子未被清除（为了避免该蛋白质对签名档的数量为 0），导致将一部分有交互的蛋白质对认为是不表达交互关系的，从而获得了较低的查全率，但是整体性能较实验方案 C_1 还是有明显的提高。实验结果说明，不同的初始化方式对本模型的迭代训练过程有很大的影响，同时基于高频词的去噪模型对训练数据的处理是合理的，使模型取得了更好的性能。

4 结束语

在基于最大期望算法的多实例多标记学习框架的基础上，对训练数据中存在的噪音问题进行处理，提出了基于高频词集的去噪模型来改进最大期望算法的初始化步骤，有效提升了蛋白质交互关系抽取模型的整体性能。下一步可以在多实例多标记模型的迭代训练过程中引入一定的先验知识，以进一步提高模型的去噪能力。

参考文献：

[1] PRASAD T S K，GOEL R，KANDASAMY K，et al. Human protein reference database－2009 update [J]. Nucleic Acids Research，2009，37：D767－D772.

[2] KERRIEN S，ALAM－FARUQUE Y，ARANDA B，et al. IntAct－open source resource for molecular interaction data [J]. Nucleic Acids Research，2007，35：D561－D565.

[3] CEOL A，ARYAMONTRI A C，LICATA L，et al. MINT，the molecular interaction database：2009 update [J]. Nucleic Acids Research，2010，38：D532－D539.

[4] BADER G D，BETEL D，HOGUE C W V. BIND：the biomolecular interaction network database [J]. Nucleic Acids Research，2003，31(1)：D248－D250.

[5] 王宇伟，牛 耘. 基于关系相似性的蛋白质交互作用识别[J]. 计算机技术与发展，2015，25(2)：42－46.

[6] BUNESCU R C，MOONEY R J. A shortest path dependency kernel for relation extraction [C]//Proceedings of the conference on human language technology and empirical methods in natural language processing. Vancouver，British Columbia，Canada：Association for Computational Linguistics，2005：724－731.

[7] AIROLA A，PYYSALO S，BJÖRNE J，et al. A graph kernel for protein－protein interaction extraction [C]//Proceedings of the workshop on current trends in biomedical natural language processing. Columbus，Ohio：Association for Computational Linguistics，2008：1－9.

[8] 唐 楠，杨志豪，林鸿飞，等. 基于多核学习的医学文献蛋

（下转第 72 页）

续表 1

手势	样本个数	识别个数	错误识别	正确率/%	拒绝识别个数	拒绝率/%	有效率/%
手势 3	16	14	0	100	2	12.5	87.5
手势 4	16	13	3	76.9	3	18.75	81.25
总计	64	55	5	90.9	9	14.1	85.9

提出的基于手势轮廓像素变化的手势识别方法计算量小,识别速度快,系统稳定,最终也取得了较高的识别率,说明了该方案的可行性。

5 结束语

文中提出了一种基于手势轮廓像素变化的手势识别方法,利用肤色的特殊性,结合 RGB 和 HSV 双颜色空间处理的肤色检测技术更加准确地分割出手势区域;然后经过手势区域判定条件去除类肤色区域和人脸区域,得到有效的手势区域;最后,基于手势轮廓像素变化完成了 4 种手势分类识别。

实验结果表明,该系统能够实时地采集手势,对从摄像头输入的四种常用静态手势进行识别,得到了较好的识别结果。由于在手势分割阶段采用的是基于颜色空间的手势分割,当背景中有大面积类肤色干扰时,分割效果将会不理想,甚至是无法分割,从而影响后续工作的进行以及最终的识别效果。因此下一步的工作方向是提高系统对类肤色干扰的适应性和鲁棒性,并且进一步提高手势识别的准确性。

参考文献:

[1] 李加力.复杂背景下的手势识别算法研究[D].厦门:厦门大学,2013.

[2] LIU Yun,ZHANG Lifeng,ZHANG Shujun.A hand gesture recognition method based on multi-feature fusion and template matching [J].Procedia Engineering,2012,29:1678-1684.

[3] 易靖国,程江华,库锡树.复杂背景下的手势识别方法[J].数字技术与应用,2016(9):50-53.

[4] 王 兵,董洪伟,张明敏,等.基于 Kinect 的动态手势识别[J].传感器与微系统,2018,37(2):143-146.

[5] 高 晨,张亚军.基于 Kinect 深度图像的指尖检测与手势识别[J].计算机系统应用,2017,26(4):192-197.

[6] 丁 毅,曹江涛,李 平,等.复杂背景下的手势识别算法研究[J].自动化技术与应用,2016,35(8):113-116.

[7] DHRUVA N,RUPANAGUDI S R,SACHIN S K,et al.Novel segmentation algorithm for hand gesture recognition [C]//IEEE international multi conference on automation computing,control,communication and compressed sensing.Kottayam,India:IEEE,2013:383-388.

[8] HASAN M M,MISHRA P K.HSV brightness factor matching for gesture recognition system [J].International Journal of Image Processing,2010,4(5):456-467.

[9] 徐战武.静态图像肤色检测研究[D].杭州:浙江大学,2006.

[10] HE Liwen,XU Yong,CHEN Yan,et al.Recent advance on mean shift tracking:a survey [J].International Journal of Image and Graphics,2013,13(3):1350012-1-1350012-29.

[11] 付潇聪,王浩平.一种基于视觉的手势识别系统[J].电子设计工程,2017,25(17):26-30.

[12] QURESHI A,MARVI M,UNAR M A,et al.Performance analysis of skin classifiers in RGB and YCbCr,color space [C]//International multi topic conference.Karachi,Pakistan:IEEE,2014:223-228.

[13] 鲁 姝.基于视觉的手势识别研究[D].天津:天津工业大学,2016.

(上接第 68 页)

白质关系抽取[J].计算机工程,2011,37(10):184-186.

[9] 刘 念,马长林,张 勇,等.基于树核的蛋白质相互作用关系提取的研究[J].华中科技大学学报:自然科学版,2013,41:232-236.

[10] NIU Y,OTASEK D,JURISICA I.Evaluation of linguistic features useful in extraction of interactions from PubMed;application to annotating known,high-throughput and predicted interactions in I2D [J].Bioinformatics,2010,26(1):111-119.

[11] 高 飞.基于 MapReduce 的蛋白质相互作用信息抽取系统的设计与实现[D].杨凌:西北农林科技大学,2016.

[12] YANG Zhihao,HONG Li,LIN Hongfei,et al.Extraction of information on protein-protein interaction from biomedical literatures using an SVM [J].CAAI Transactions on Intelligent Systems,2008,3(4):361-369.

[13] 刘敏捷.基于组合学习和主动学习的蛋白质关系抽取[D].大连:大连理工大学,2015.

[14] ZHOU Zhihua,ZHANG Minling,HUANG Shengjun,et al.Multi-instance multi-label learning [J].Artificial Intelligence,2011,176(1):2291-2320.

[15] ZHOU Zhihua,ZHANG Minling.Multi-instance multi-label learning with application to scene classification [C]//International conference on neural information processing systems.Canada:MIT Press,2007:1609-1616.