

# 基于最大期望算法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:**针对基于远监督的方法中训练数据存在噪音的问题,采用了一种基于最大期望(EM)算法的多实例多标记的方法来进行蛋白质关系的抽取。首先通过对大规模生物医学文本的自动搜索建立目标蛋白质对的签名档,提取出签名档中的词法和语法等特征,作为蛋白质对签名档的向量空间模型(VSM);然后引入隐变量,将蛋白质对的签名档及其标签构建为多实例多标记学习模型,利用最大期望算法来迭代消除训练数据中的噪音;最后通过有监督的方法来预测未知蛋白质对的交互关系。针对蛋白质对描述中还存在的其他蛋白质名称会对交互关系的判断产生影响,改进了蛋白质对的特征表示。实验结果表明,该方法较原始的最大期望算法取得了更高且均衡的精确度和召回率。

**关键词:**蛋白质交互;最大期望算法;多实例多标记;蛋白质实体识别

**中图分类号:**TP391

**文献标识码:**A

**文章编号:**1673-629X(2018)08-0048-05

doi:10.3969/j.issn.1673-629X.2018.08.010

## Protein-protein Interaction Identification Based on Expectation Maximization Algorithm

CAI Song-cheng, NIU Yun

(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,  
Nanjing 211106, China)

**Abstract:** In order to solve the problem of noise in training data based on remote supervision, a multi-instance multi-label method based on maximum expectation (EM) algorithm is adopted to extract protein relations. The signature of a protein pair is obtained first by searching large scale biomedical text and lexical and syntactic features are extracted to form protein pair's vector space model (VSM). Then, we jointly model the signatures of protein pairs and their labels using MIML learning with latent variable and reduce noise iteratively by using EM algorithm. Finally, we predict whether unknown protein pairs are interactive or not with supervised method. As the signature of the target protein pair usually contains other proteins which may affect the judgment of the interaction between target protein pairs, we improve the feature expression of protein pairs. The experiment shows that the method has achieved high and well balanced precision and recall compared to the original EM algorithm.

**Key words:** protein-protein interaction; expectation maximization algorithm; multi-instance multi-label; protein entity recognition

## 0 引言

随着人们对文本中分子途径和分子交互关系等信息需求的不断增加,蛋白质交互作用关系(protein-protein interaction, PPI)的自动抽取在分子生物学领域变得越来越重要。PPI是指细胞内两个蛋白质之间的交互作用,这种交互作用环环相扣,深刻影响着整个细胞生理作用的调节。起初生物医学领域的专家手工地从医学文献中收集这些信息录入统一格式的数据库中,如 HPRD<sup>[1]</sup>、IntAc<sup>[2]</sup>、MINT<sup>[3]</sup>和 BIND<sup>[4]</sup>等。然而随着生物医学文献的急剧增加,新的蛋白质之间的关

系也在产生。手工录入蛋白质之间的交互信息显然远不能满足实际需要,因此自动地从医学文献中抽取 PPI 已经成为一项重要的研究内容。

在此背景下,基于自然语言处理的 PPI 自动识别技术正在快速发展并已取得很大的进展。目前 PPI 识别是采用有监督的机器学习方法,以单句为依据来识别句子之间的交互关系,需要大量人工标注的数据,代价高昂,所以将远监督的思想运用到 PPI 识别上,解决了训练数据不足的问题。但是由于远监督思想的缺陷,引入了大量噪音,影响现阶段 PPI 识别的精度。针

收稿日期:2017-09-29

修回日期:2018-01-11

网络出版时间:2018-04-28

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://www.cnki.net/kcms/detail/61.1450.TP.20180427.1626.016.html>

对这个问题,采用一种基于最大期望算法的多实例多标记学习(multi-instance multi-label, MIML)方法来进行蛋白质交互关系的识别,有效消除了签名档中噪音对交互关系识别的影响。

## 1 相关工作

目前,用于从生物医学文献中抽取 PPI 的技术主要包括:基于同现的方法<sup>[5]</sup>、基于规则的方法和基于机器学习<sup>[6-8]</sup>的方法。基于同现的方法通过统计两个蛋白质在句子中的共现频率来判断是否存在交互关系,识别结果召回率高但精确度低;基于规则的方法可以取得较高的精确度但是召回率较低,而且通过手动建立规则的方法需要大量的人力物力,且制定的规则只适用于某些特定领域的的数据,无法普遍应用。

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 的识别。基于机器学习的方法主要包括两大类:基于特征的方法和基于核函数的方法。基于特征的方法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<sup>[9-10]</sup>。基于核函数的方法首先深入研究句子结构,通过设计核函数进一步利用句子结构表示(如字符串序列、句法依赖或句法分析)上的隐含特征,然后使用支持核函数的分类器进行 PPI 关系的识别。Haussler D<sup>[11]</sup>提出了针对离散结构的卷积核;Lodhi H 等<sup>[12]</sup>将特征空间特定长度词语子序列的内积作为函数的计算方式,提出了字符串核;Bunescu R C 等<sup>[13]</sup>提出了最短依赖路径核,将句子以树的形式表示,用两个实体之间的最短路径表示实体之间的关系。然而目前利用机器学习方法来进行 PPI 关系识别一般都是以句子为单位,分析一句话中出现的任意一对蛋白质对之间是否存在交互关系。这种方式能够在句子级别上提供蛋白质对交互关系的描述和证据,但是也存在一定的局限性。这种方式所需的训练集要求对每一个句子中出现的每一对蛋白质是否存在交互关系进行标注,当训练语料不足时,PPI 关系识别的效果会大打折扣。但人工标注大规模文本需要耗费大量的人力物力。

针对这些不足,文中试图采用远监督思想来进行 PPI 关系的抽取。远监督方法已经用于关系识别领域,远监督思想假设如果两个实体之间存在某种关系,那么包含这两个实体的所有句子都在一定程度上表达了这种关系。基于上述假设,远监督通过将知识库中的实体和训练语料文本中的实体进行匹配,产生大量带标注的训练数据,避免了人工标注数据的繁重劳动。对于 PPI 关系识别,同样存在标注数据不足的问题,所以可以将远监督<sup>[15]</sup>方法运用到 PPI 关系抽取上。

但是基于远监督方法的 PPI 识别也存在一个问题。对于有交互关系的蛋白质对事实上并非其签名档中的所有句子都表达了该蛋白质对的交互关系,其中很多句子是不表达交互关系的,从而这部分数据成为了训练过程中的噪音,最终会影响蛋白质对交互关系的识别结果。

针对远监督的 PPI 抽取方法存在的问题,文中采用一种基于最大期望算法的多实例多标记的学习方法。多实例多标记是一种新型的关系抽取的学习框架<sup>[16]</sup>,在该框架中,每个对象由多个实例描述,同时对象可以拥有多个类别标记,这个框架尤其适用于多义性的对象。多实例多标记学习框架已被成功应用于图像文本分类<sup>[17]</sup>、视频标注<sup>[18]</sup>、基因图像识别<sup>[19]</sup>等任务中,既充分利用了蛋白质对签名档的信息,同时又改善了利用远监督思想来标记签名档中的句子带来的噪音问题。在此基础上又对特征加以改进,有效消除了其他蛋白质对目标蛋白质对交互关系识别的影响。

## 2 基于最大期望算法的 PPI 识别

基于最大期望算法的多实例多标记学习方法,是在基于远监督方法的基础上,从大规模生物医学文献中搜索得到的蛋白质对签名档中提取特征,构建向量空间模型(vector space model, VSM)。在此基础上引入隐变量,将蛋白质对的签名档和标签构建为多实例多标记的学习框架,利用最大期望算法迭代地消除噪音。最终采用监督学习的方法来预测未知蛋白质对的交互关系。

### 2.1 关系提取

PubMed 数据库作为建立 PPI 网络重要的数据来源,收录了超过一千八百万篇生物医学文献摘要。从 PubMed 数据中获取蛋白质对签名档的过程包括:

(1)调用 PubMed 数据库提供的接口,搜索包含目标蛋白质对的摘要。

(2)使用伊利诺州大学 Urbana-Champaign 分校认知计算研究组开发的句子识别工具来识别摘要集合中的句子,保留包含目标蛋白质对的句子作为签名档的内容。

最终每一个目标蛋白质对都会有一个包含多个句子的集合与之对应,这个句子集合即为蛋白质对的签名档,接下来将签名档作为蛋白质对交互关系的特征来源进行处理。

### 2.2 特征表示

实验中使用到了两个逻辑回归分类器来进行 PPI 关系的识别。一个是对蛋白质对签名档中的句子进行交互关系判断的句子级分类器,另一个是对蛋白质对进行分类的顶层分类器。两个分类器的主要差别在于

特征的表示上,句子级分类器利用提取得到的句子的语言学特征进行分类,而顶层分类器通过当前签名档中句子的分类结果形成特征进行分类。句子级分类器特征的形成主要是选取训练集中所有句子中重要的单词特征作为向量的每一维。具体处理过程为:首先对句子进行分词,去除无意义的标点符号以及停用词;然后选取句子中出现在两个目标蛋白质之间的单词,以及第一个目标蛋白质左边 2 个单词和第二个目标蛋白质右边 2 个单词;最终将这些单词作为句子中蛋白质对的上下文特征来构建向量空间模型。若在句子的上下文特征中出现了某个特征词,则在向量中对应于出现特征词的某一维用 1 记录,否则用 0 记录。

对于顶层分类器中蛋白质对的实际交互关系,采取签名档中判断为有交互关系的句子数占签名档中所有句子的比例作为特征构建一维向量。

### 2.3 多实例多标记学习模型

在该模型中,对于训练集中的每一个蛋白质对,都有已知的唯一标记,即有无交互关系,但对于签名档中的每一个句子并不知道其真实的标记。所以,引入一个隐变量  $z$  来代表句子的标记。 $z = \text{non-interactive}$  表示在该句中目标蛋白质对之间没有交互关系; $z = \text{interactive}$  表示目标蛋白质对之间存在交互关系。对于 PPI 关系抽取中的关系是互补的,两个蛋白质之间的关系就分为有交互和无交互两种。在该模型中,如图 1 所示,由两层构成,包含一个对蛋白质对签名档中的句子进行分类的句子级二元分类器( $z$  分类器)和一个对蛋白质对进行分类的顶层二元分类器( $y$  分类器)。

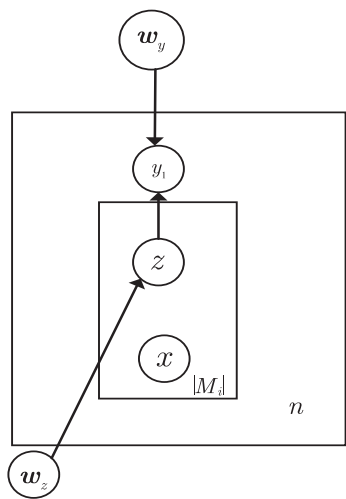


图 1 多实例多标记学习框架

图中,  $n$  表示蛋白质对的数目;  $M_i$  表示第  $i$  对蛋白质对签名档的数目;  $x$  表示输入的一个句子;  $w_z$  表示  $z$  分类器的权重向量;  $w_y$  表示  $y$  分类器的权重向量。

#### 2.3.1 训练

由于蛋白质对签名档中的句子标记是未知的,而最大期望算法是估计隐变量的有效方法,所以文中采

用最大期望算法来训练多实例多标记的学习框架。最大期望算法主要由  $M$  步和  $E$  步构成,  $M$  步训练句子级分类器( $z$  分类器)和顶层分类器( $y$  分类器),  $E$  步根据得到的两个分类器来更新句子的标记,经过多次迭代使句子的标记更加接近于真实的标记。

在以下的公式中,向量  $z_i$  代表第  $i$  个蛋白质对所有句子的标记构成的一个向量;  $y_i$  用来表示第  $i$  个蛋白质对的标记,用公式表示为:

$$y_i^{(r)} = \begin{cases} 1, & r \in P_i \\ 0, & r \in N_i \end{cases} \quad (1)$$

其中,  $P_i$  为关系正例,表示第  $i$  对蛋白质对具有的关系;  $N_i$  是关系负例,表示第  $i$  对蛋白质对不具有的关系。

文中使用最大期望算法来最大化极大似然函数的下界,也就是说最大化数据库中每个蛋白质对的联合概率,得到:

$$\log p(y_i, z_i | x_i, w_y, w_z) = \sum_{m \in M_i} \log p(z_i^{(m)} | x_i^{(m)}, w_z) + \sum_{r \in P_i \cup N_i} \log p(y_i^{(r)} | z_i, w_y^{(r)}) \quad (2)$$

**E-step:** 在此步骤,对于每个蛋白质对的签名档,给定蛋白质对的标记集合,以及目前模型学习得到的  $z$  分类器和  $y$  分类器的权重向量,推断出蛋白质对句子级别的分类结果。

$$z_i^* = \operatorname{argmax}_z p(z | y_i, x_i, w_y, w_z) \quad (3)$$

通过近似化,将向量  $z$  进行拆分,分开考虑每个句子的分类结果。对于每个蛋白质对  $i = 1, 2, \dots, n$  中的每个句子  $m \in M_i$ , 计算:

$$z_i^{(m)*} = \operatorname{argmax}_z p(z | x_i^{(m)}, w_z) \times \prod_{r \in P_i \cup N_i} p(y_i^{(r)} | z_i', w_y^{(r)}) \quad (4)$$

其中,  $z_i'$  表示上一次迭代得到的第  $i$  个蛋白质对的句子分类结果,除了句子  $m$  其句子标记用  $z_i^{(m)*}$  替代。

**M-step:** 此步骤利用 **E-step** 得到的句子分类结果  $z_i$  通过最大化似然函数的下界,得到对应的  $w_z$  和  $w_y$ 。实际上就是通过学习来更新句子级和蛋白质对级权重参数,具体公式如下:

$$w_z^* = \operatorname{argmax}_w \sum_{i=1}^n \sum_{m \in M_i} \log p(z_i^{(m)*} | x_i^{(m)}, w) \quad (5)$$

$$w_y^{(r)*} = \operatorname{argmax}_w \sum_{1 \leq i \leq n \text{ s.t. } r \in P_i \cup N_i} \log p(y_i^{(r)} | z_i^*, w) \quad (6)$$

#### 2.3.2 预测

(1) 对于一个给定的蛋白质对,首先预测其签名档中句子的分类结果。

$$z_i^{(m)*} = \operatorname{argmax}_z p(z | x_i^{(m)}, w_z) \quad (7)$$

(2) 利用顶层分类器来决定该蛋白质对是否具有交互关系。



$$y_i^{(r)*} = \operatorname{argmax}_{y \in \{0,1\}} p(y | z_i^*, w_y^{(r)}) \tag{8}$$

2.3.3 实 现

初始化:由于最大期望算法并不是全局最优算法,因此初始值的设置对最后的结果有着重要的影响。在该模型中,初始值为签名档中句子的类别分布  $z_i$ 。利用原始的签名档数据来训练一个分类器,然后通过此分类器对签名档中的句子进行分类,将分类结果作为初始值  $z_i$ 。

2.4 特征改进

通过对目标蛋白质对签名档数据的观察,发现在包含目标蛋白质对的同一个句子的描述中往往还存在其他蛋白质,这些蛋白质可能会对目标蛋白质交互关系的判断造成影响。基于这个原因,需要对句子级分类器原始的特征加以改进。

文中利用一个生物医学文本命名实体识别工具 ABNER 来识别句子中其他蛋白质的名称。ABNER 在 NLPBA 和 BioCreative 语料库上进行训练,在两个语料库上识别的  $F$  值分别达到了 72.6% 和 69.9%。

通过观察蛋白质对的签名档,可以发现描述交互作用的句子中经常会出现 bind、interact、activate、inhibit、down-regulate 等表示蛋白质交互作用的单词。这些单词通常被认为是识别蛋白质交互关系的关键词。关键词对于蛋白质交互关系识别尤为重要,已经作为线索运用到基于模式匹配的 PPI 抽取方法中。文中选择关键词作为一维特征对原有特征加以改进,采用的关键词集合利用了 Joshua M. Temkin<sup>[20]</sup> 提出的关键词列表。

观察以下描述蛋白质交互关系的句子:

#arnt# mRNA appeared to be slightly but significantly down-regulated by <protein> BaP </protein> as well as by flavonoids while expression of #aip# was not or only slightly modulated.

用##标注出来的是两个目标蛋白质,而用<protein></protein>标注的是利用 ABNER 工具识别出来的其他蛋白质,蛋白质全部由黑体显示。在这个句子中,有一个关键词“down-regulated”的出现,很有可能说明目标蛋白质对(arnt,aip)之间存在交互关系,但由于在目标蛋白质之间出现了一个其他蛋白质 BaP,这个关键词事实上表达出来的是其他蛋白质 BaP 和目标蛋白质 arnt 之间的交互关系,所以有必要将包含目标蛋白质对的同一个句子中的其他蛋白质识别出来。

在保留 2.1 节所有特征的基础上,又新增了 5 个特征来对句子级分类器进行改进。首先对签名档中的每一个句子,抽取出第一个蛋白质左边的 7 个单词和第二个蛋白质右边的 7 个单词以及两个蛋白质中间的所有单词。然后,将第一个蛋白质左边和第二个蛋白质

右边是否有关键字和其他蛋白质的名称以及目标蛋白质中间有没有其他蛋白质作为 5 维特征添加到原有特征中,权重采用二值权重,若有则置为 1,否则置为 0。增加了这 5 维特征后,以第一个蛋白质左边的两个特征为例,若出现了关键词和其他蛋白质,则很有可能表示的是其他蛋白质和第一个目标蛋白质之间的交互关系。

3 实 验

3.1 实验数据及设置

采用的训练数据来自于现有的 PPI 数据库,无需额外的人工标注。将有交互关系的蛋白质对视为正样例,无交互的视为负样例。实验中有交互关系的蛋白质对是直接 from 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,采用生物医学领域常用方法,将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。以两个待考察的蛋白质为查询条件,通过 PubMed 数据库的应用程序接口查询目标蛋白质对的文献摘要,然后对摘要文本集合进行处理,找出包含目标蛋白质对的句子,形成签名档。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。

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 × 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。将原始数据按照蛋白质对平均划分为 5 折,将每个子集数据分别做一次验证集,其余的 4 组子集数据作为训练集,这样会得到 5 个模型,用这 5 个模型最终验证集的平均性能作为评价整个方法性能的指标。

3.2 实验结果及分析

为了比较使用原始特征和改进后特征的实验结果,以第一折数据为例,取最大期望算法迭代的前六次 (迭代 6 次以后实验结果基本趋向局部最优解),结果如表 1、表 2 所示。

表 1 采用原始特征的识别结果

| 迭代次数 | $P$  | $R$  | $F$  |
|------|------|------|------|
| 1    | 0.75 | 0.66 | 0.7  |
| 2    | 0.74 | 0.68 | 0.71 |
| 3    | 0.80 | 0.70 | 0.75 |
| 4    | 0.78 | 0.73 | 0.76 |
| 5    | 0.72 | 0.73 | 0.73 |
| 6    | 0.71 | 0.75 | 0.73 |

表 2 采用改进特征的识别结果

| 迭代次数 | <i>P</i> | <i>R</i> | <i>F</i> |
|------|----------|----------|----------|
| 1    | 0.75     | 0.68     | 0.71     |
| 2    | 0.74     | 0.68     | 0.71     |
| 3    | 0.75     | 0.75     | 0.75     |
| 4    | 0.79     | 0.78     | 0.78     |
| 5    | 0.77     | 0.76     | 0.77     |
| 6    | 0.72     | 0.77     | 0.74     |

从这两张表可以发现,随着迭代次数的增加,采用改进以后的特征在精确度、召回率和 *F* 值上都有明显提高。最终结果是要把五折数据识别的平均结果作为该模型 PPI 识别的性能,如表 3 所示。

表 3 五折交叉验证识别结果比较

| 特征   | 精确度   | 召回率   | <i>F</i> 值 |
|------|-------|-------|------------|
| 原始特征 | 0.688 | 0.74  | 0.712      |
| 改进特征 | 0.676 | 0.776 | 0.722      |

从上述识别结果发现,对特征加以改进后,识别的准确率虽然稍有下降,但是召回率提高了 3.6%,整体 *F* 值提高 1%。说明改进后,算法考虑了其他蛋白质对目标蛋白质识别的影响,使模型取得了更好的性能。

4 结束语

由于基于远监督的 PPI 抽取方法存在大量噪音问题,文中采用基于最大期望算法的多实例多标记学习框架,同时在此基础上对特征加以改进,消除了签名档中其他蛋白质对目标蛋白质对交互关系判断的影响。实验结果表明,该方法取得了更高的识别精度。

下一步将利用蛋白质对签名档中包含的丰富信息对句子级分类器得到的结果进行改进,使句子级的分类更加准确,从而能进一步提高 PPI 识别的效果。

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,2009,37:767-772.

[2] KERRIEN S,ALAM-FARUQUE Y,ARANDA B,et al. IntAct-open source resource for molecular interaction data[J]. Nucleic Acids Research,2007,35:561-565.

[3] CEOL A,ARYAMONTRI A C,LICATA L,et al. MINT,the molecular interaction database:2009 update[J]. Nucleic Acids Research,2010,38:532-539.

[4] BADER G D,BETEL D,HOGUE C W V. BIND;the bio-molecular interaction network database[J]. Nucleic Acids Research,2003,31(1):242-245.

[5] BUNESCU R,MOONEY R,RAMANI A,et al. Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline[C]//Proceedings of the workshop on linking natural language pro-

cessing and biology:towards deeper biological literature analysis. [s. l.]: Association for Computational Linguistics, 2006:49-56.

[6] 杨志豪,洪莉,林鸿飞,等. 基于支持向量机的生物医学文献蛋白质关系抽取[J]. 智能系统学报,2008,3(4):361-369.

[7] 崔宝今,林鸿飞,张霄. 基于半监督学习的蛋白质关系抽取研究[J]. 山东大学学报:工学版,2009,39(3):16-21.

[8] 董美豪. 基于文本挖掘的蛋白质相互作用对抽取方法的研究[D]. 哈尔滨:哈尔滨工业大学,2015.

[9] 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.

[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] HAUSSLER D. Convolution kernels on discrete structures[R]. California: University of California at Santa Cruz, 1999.

[12] LODHI H,SAUNDERS C,SHAWE-TAYLOR J,et al. Text classification using string kernels[J]. Journal of Machine Learning Research,2002,2(3):419-444.

[13] 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. [s. l.]: Association for Computational Linguistics,2005:724-731.

[14] 封二英,牛耘,魏欧. 基于大规模文本的蛋白质交互关系自动提取[J]. 计算机应用,2012,32(S1):147-150.

[15] 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.

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

[17] 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.

[18] XU Xinshun,JIANG Yuan,XUE Xiangyang,et al. Semi-supervised multi-instance multi-label learning for video annotation task[C]//Proceedings of the 20th ACM international conference on multimedia. Nara, Japan: ACM,2012:737-740.

[19] LI Yingxin,JI Shuiwang,KUMAR S,et al. Drosophila gene expression pattern annotation through multi-instance multi-label learning[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics,2012,9(1):98-112.

[20] TEMKIN J M,GILDER M R. Extraction of protein interaction information from unstructured text using a context-free grammar[J]. Bioinformatics,2003,19(16):2046-2053.