高技术通讯 2009 年 第 19 卷 第 5 期:480 ~ 486

doi:10.3772/j.issn.1002-0470.2009.05.008

基于 TSVM 与主动学习融合的蛋白质交互作用关系抽取①

刘健苗②* ** 王浩畅* *** 赵铁军*

(*哈尔滨工业大学教育部-微软语言语音重点实验室 哈尔滨 150001)
(**中讯邮电咨询设计院有限公司信息工程处 郑州 450007)
(***大庆石油学院计算机与信息技术学院 大庆 163318)

摘 要 针对蛋白质交互作用关系(PPI)抽取研究中已标注语料有限而未标注生物医学自由文本易得的问题,进行了基于直推式支持向量机(TSVM)与主动学习融合的蛋白质交互作用关系抽取研究。通过自主选择最优的未标注样本加入到 TSVM 的训练过程中,最大程度地提高了系统的性能。实验结果表明,TSVM 与主动学习融合的算法在少量已标注样本和大量未标注样本组成的混合样本集上取得了较好的学习效果,与传统的支持向量机(SVM)和 TSVM 算法相比,能有效地减少学习样本数,提高分类精度,在 AImed 语料上取得了 F 测度为 64.12% 的较好性能。

关键词 蛋白质交互作用关系抽取,半监督学习,直推式支持向量机(TSVM),主动学习

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。所谓 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,从而形成了一个巨大的网状关系,深刻地影响着整个细胞生理作用的调节。PPI 抽取旨在从生物医学文献中识别蛋白质、药物或其他分子间的转录、翻译、翻译后修饰、络合和分解等关系[1]。当前,PPI 抽取主要采用有监督学习的方法,并且已取得一定成绩[1-4]。利用大规模的已标注语料固然可以提高分类系统的性能,然而已标注语料的获得需要领域专家付出大量的精力手工完成,这显然跟不上当代信息增长的步伐。因此,迫切需要一种在小规模标注语料上同样能得到很好分类结果的高性能学习方法。半监督学习作为解决这类问题的一种方法应运而生,它在传统的有监督学习中结合未标注样本和已标注样本进行学习,通过它们的联合概率分布来改进分类器的性能,逐渐成为研究的热点。

直推式支持向量机(transductive support vector machine, TSVM)[5]作为一种新颖的半监督学习方法,已经被应用于文本分类、图像分类、生物技术等领域[5-8]。TSVM 假定未标注样本就是测试样本,学习的目的就是在这些未标注样本上取得最佳泛化能力。训练过程中以少量的已标注样本和大量的未标注样本进行学习,使得未标注样本的样本分布信息转移到最终的分类器中。未标注样本规模较大,且与已标注样本独立同分布,所以能够更好地刻画整个样本空间的数据特性,从而使训练出的分类器具有更好的分类性能。由于已标注样本的成本太高,我们希望将海量的自由文本作为未标注样本加入到 TSVM 的学习过程中。然而,判断来自不同环境的自由文本是否含有噪声,是否与已标注样本同分布是一件非常困难的工作。本文在对 TSVM 和主动学习进行研究的基础上,提出了一种 TSVM 与主动学习融合的 PPI 抽取学习算法,采用主动学习的方法对这些自由文本进行指导性地选择。实验结果表明,我们的方法能够使分类器在小规模已标注样本环境中具有较高的分类精度。

1 研究方法

1.1 直推式支持向量机(TSVM)

传统的归纳式支持向量机需要大量已标注样本

① 863 计划(2006AA01Z150)和国家自然科学基金(60736044)资助项目。
② 男,1980 年生,硕士,助教;研究方向:自然语言处理,信息抽取;联系人,E-mail:liujianmiao@yahoo.cn
(收稿日期:2008-06-26)

来训练分类器,而样本的正确标注需要手工完成,代价比较昂贵。生物医学领域中常用于蛋白质交互作用关系抽取的 AImed 评测语料也只标注了 1000 多个蛋白质关系,其规模很难满足实际需要。如果能把容易获取的未标注样本加入到 PPI 抽取过程中,就可以弥补归纳式支持向量机的缺陷,TSVM 算法正是基于这种思想的支持向量机(support vector machine, SVM)算法[9]。作为标准支持向量机算法在直推式学习问题上的一种扩展,TSVM 使用未标注样本的内部特征以加大支持向量机最优分类超平面的分类间隔(margin),较大的分类间隔就意味着较好的分类泛化能力。下面简单介绍 TSVM 算法的原理,详细地描述和证明参见文献[4]。

给定一组已标注训练样本 $(x_1, y_1), \cdots, (x_n, y_n)$, $x_i \in R^m, y_i \in \{-1, +1\}$ 和另一组具有相同分布的未标注样本 $x_1^*, \cdots, x_k^*$, TSVM 的训练过程就是为了寻找一种最优的求解方法来确定 $x_1^*, \cdots, x_k^*$ 的待分类值 $y_1^*, \cdots, y_k^*$, 使得分类超平面 $w \cdot x + b = 0$ 在联合序列 $(x_1, y_1), \cdots, (x_n, y_n), (x_1^*, y_1^*), \cdots, (x_k^*, y_k^*)$ 上的分类间隔最大化。

在一般线性不可分条件下,可以将上述问题形式化描述为如下最优化问题:

Minimize over $(y_1^*, \cdots, y_k^*, w, b, \xi_1, \cdots, \xi_n, \xi_1^*, \cdots, \xi_k^*)$

$$\frac{1}{2}\| w \|^2 + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{k}\xi_j^*$$

$$\text{subject to: } \forall_{i=1}^{n}: y_i[w \cdot x_i + b] \geqslant 1 - \xi_i$$

$$\forall_{j=1}^{k}: y_j[w \cdot x_j^* + b] \geqslant 1 - \xi_j^*$$

$$\forall_{i=1}^{n}: \xi_i \geqslant 0$$

$$\forall_{j=1}^{k}: \xi_j^* \geqslant 0 \qquad (1)$$

其中参数 C 和 C^* 是用户指定和调节的参数,分别为已标注样本和未标注样本在训练过程中的影响因子, ξ_i 和 ξ_j^* 分别为已标注样本和未标注样本在训练过程中被错分的惩罚因子。TSVM 的训练过程如下:(1)指定参数 C 和 C^*, 对已标注样本进行归纳式学习,得到初始分类器,并指定未标注样本中正例的个数 N;(2)用初始分类器对未标注样本分类,对输出值最大的 N 个未标注样本暂赋正值,其余的赋负值,并指定临时影响因子 C_{tmp}^*; (3)对所有样本重新训练以得到新分类器,按一定规则交换一对标签值不同的未标注样本的标签符号,使优化问题式 Minimize over $(y_1^*, \ldots, y_k^*, w, b, \xi_1, \ldots, \xi_n, \xi_1^*, \ldots, \xi_k^*)$ 中的目标函数值下降最大,反复执行这一步直到找不出符合交换条件的样本对为止;(4)均匀地增加 C_{tmp}^* 的值并返回到步骤(3),当 $C_{tmp}^* \geqslant C^*$ 时算法结束,输出所有未标注样本的标签值。

1.2 TSVM 与主动学习融合算法

主动学习[10]是指使用机器学习方法迭代地从候选样本集中按某种策略动态选择样本进行训练的过程。由于分类器每次能从大量的未标注样本中识别出一些具有高训练效用的样本子集,因而不需要花费大量时间去标注那些对进一步学习没有帮助或者帮助不大的样本。另一方面,由于控制了训练样本的大小,训练分类器的计算规模也能够得以缩减。

主动学习是一个循环反复的标注学习过程。首先,根据先验知识或者随机地从候选未标注样本集中选择少量样本并正确标注类别,构造至少包含一个正例和一个负例的初始训练样本集。然后,通过使用该初始训练样本集训练得到的初始分类器,采用某种样本选取算法,从候选未标注样本集中选择最有利于提高分类器性能的样本,正确标注类别后加入到训练样本集中,重新训练分类器,再次选择最有利于提高分类器性能的样本。重复以上过程直到候选样本集为空或达到某种指标[11]。

TSVM 与主动学习融合的基本思想是从候选未标注样本集中选取最不确定的样本和从测试样本中选取最确定的样本加入到学习过程中,从而对分类超平面进行调整。因此,学习过程中的样本选取策略非常关键,直接关系到整个算法的性能。受文献[12]启发,我们认为 SVM 算法的分类绝对值可以使基于样本的不确定性方法与版本空间和边缘方法这两个主动学习策略达到统一。即离分类超平面越近的样本分类绝对值越小,越可能是最不确定的样本;反之,离分类超平面越远的样本分类绝对值越大,其分布特征更显著,分类正确的可能性和样本的确定性也越大,同时也越不可能是噪声。因此,本文提出以下两种样本选取策略:在每次迭代过程中,(1)从候选未标注样本集中选取最不确定的若干个样本并进行正确标注;(2)从测试样本集中选取最确定的若干个样本。基于 TSVM 与主动学习方法融合的蛋白质交互作用关系抽取算法流程描述如下。

算法由 SVM 训练和 TSVM 迭代两个部分组成。首先,从候选未标注样本集 S_{untag} 中根据先验知识或者随机地选择少量样本并正确标注它们的类别,构造初始训练样本集 S_{train},确保 S_{train} 中至少包含有一个正例和一个负例;然后,利用 S_{train} 训练得到的 SVM 分类器 C 对测试样本集 S_{test} 进行标注,使用样

本选取策略 2 从 S_{test} 中抽取所有最确定的样本，加入到 S_{train} 中组成新的训练集 S^*_{train} 训练 TSVM 分类器 C*；最后用分类器 C* 对剩下的未标注样本集与全部测试样本集重新标注，调用样本选取策略 1 和策略 2 构建新的训练集 S^*_{train}。反复迭代直至达到算法的最大迭代次数。实验中，以 0.5 作为判断样本确定性的阈值，分类绝对值大于 0.5 的样本是最确定的样本，否则不是。Tong 等[13]指出在样本选取策略 1 中一次标注 1 个样本可以获得最好的分类性能，但由此需要学习的次数就越多，计算复杂度就越大，实验中我们一次标注 5 个样本。TSVM 与主动学习融合算法（TSVM _ AL 算法）步骤如表 1 所示。

表 1 TSVM 与主动学习融合算法

TSVM _ AL (S_{test}, S_{untag}, MaxIter, i, n)
输入参数： S_{test}：测试样本集； S_{untag}：候选未标注样本集； *MaxIter*：TSVM 训练最大迭代次数； i：从 S_{untag} 中选取用于训练初始 SVM 分类器 C 的样本数； n：样本选取策略 1 中选取的最不确定样本数。
算法步骤： 1) 从候选未标注样本集 S_{untag} 中随机选取 i 个样本并正确标注，其中至少包含一个正例和一个负例，作为初始训练样本集 S_{train} 训练 SVM 分类器 C； 2) 利用分类器 C 对 S_{test} 中的样本进行分类，按照样本选取策略 2 从 S_{test} 中选取分类绝对值大于阈值的样本，构建 S^*_{train} 样本集； 3) Iter ← 1, while ($Iter \leqslant MaxIter$) a) 以 S_{train} 为已标注样本，S^*_{train} 为未标注样本训练 TSVM 分类器 C^*； b) 利用分类器 C^* 分别对样本集 S_{test} 和 $S_{untag} - S_{train}$ 进行分类，得到各自的分类值； c) 从 S_{test} 中按样本选取策略 2 选择分类绝对值大于阈值的样本重新构建 S^*_{train}； d) 从 $S_{untag} - S_{train}$ 中按样本选取策略 1 选择分类绝对值最接近 0 的 n 个样本，正确标注后加入到 S_{train} 中； e) $Iter \leftarrow Iter + 1$； 4) 迭代结束，用最后得到的 TSVM 分类器 C^* 对样本集 S_{test} 进行分类。
输出结果： S_{test} 的分类值 $y^*_1, \ldots, y^*_k$。

2 特征选择

特征选择在机器学习方法中起着至关重要的作用。特征选择的主要目的是寻找那些能够帮助抽取蛋白质交互作用关系的文本相关属性，一个好的特征将具有较强的区分度。在我们的研究中，主要使用了以下浅层语言学特征：

（1）候选蛋白质实体对词特征。包括出现在两个候选蛋白质名称中的所有单词。为了减少数据的稀疏性，我们使用缩写词识别算法识别语料中的候选蛋白质名称缩写形式，然后将每个候选蛋白质名称映射到它唯一的缩写形式。如果候选蛋白质名称没有可获取的缩写形式时，则使用其原形。

（2）上下文特征。两个候选蛋白质名称周围的词。根据这些词的位置不同，将它们分为 3 个部分：两个候选蛋白质名称中间的词、第一个候选蛋白质名称左边的词和第二个候选蛋白质名称右边的词。研究发现两个候选蛋白质名称之间的词汇特征对关系抽取有很大作用。另外，在语法结构中有倒装等其他的语法结构，所以第一个候选蛋白质名称左边的词和第二个候选蛋白质名称右边的词也有可能成为判断关系类别的关键信息。但是，如果把第一个候选蛋白质名称左边的词和第二个候选蛋白质名称右边的词都包含进来又可能会引入噪声。实验研究发现，一般引入第一个候选蛋白质名称左边的 2-3 个词和第二个候选蛋白质名称右边的 2-3 个词要比引入更多或更少的词效果更好。因此，我们选择的特征包括两个候选蛋白质名称之间的所有词，如果两个蛋白质名称之间没有任何词，特征值设为 NULL；两个候选蛋白质名称周围的词，包括第一个候选蛋白质名称左边的 3 个词和第二个候选蛋白质名称右边的 3 个词，如果没有词，则特征值设为 NULL。所有词看作 bag-of-word，即不考虑词的顺序。

（3）关键词特征。表示蛋白质交互作用关系常用的一些词，通常是动词。如果两个候选蛋白质名称之间或者周围存在关键词，则关键词、关键词的词性及其位置作为特征。位置特征分为三类：分别指关键词在两个候选蛋白质名称之间，在第一个候选蛋白质名称左边 3 个词中，或者在第二个候选蛋白质名称右边 3 个词中。关键词距最近的候选蛋白质名称的距离用词的个数表示。

（4）语块特征。包括以下三个特征集合：1）出现在两个候选蛋白质名称之间的语块中心词，与上

下文特征类似,这些语块中心词也不考虑词的顺序。2)候选蛋白质名称对周围所有的语块中心词。包括第一个候选蛋白质名称左边两个语块的中心词和第二个候选蛋白质名称右边一个语块中心词。3)两个候选蛋白质名称之间所有的语块类型,用"_"组合成一个特征。

(5) 两个候选蛋白质名称的中心词对特征。我们选取两个候选蛋白质名称的一元中心词对和二元中心词对作为特征。由于英语名词短语一般为右分叉结构,其最后一个词多为中心词,因此我们选择两个候选蛋白质名称各自最后一个词组成一元中心词对特征,各自最后两个词组成二元中心词对特征。如果最后一个词是数字、希腊字符或者罗马字符及一些不能作为中心词的特殊词,那么选倒数第二个。选取中心词作为特征能够减少数据的稀疏问题。

(6) 两个候选蛋白质名称中心词的后缀对和前缀对特征。包括一元中心词后缀对特征和前缀对特征。我们用统计的方法从训练语料中统计出频率较高的蛋白质名称前后缀词表,当候选蛋白质名称包含词表中的前后缀时,则使用此表中的前后缀特征,否则取候选蛋白质名称的前三个字符作为前缀特征和后三个字符作为后缀特征。

(7) 句子中其他蛋白质名称特征。出现在两个候选蛋白质名称之间的其它蛋白质名称的数量。一个句子中有时包含多个蛋白质名称,相邻的蛋白质更可能具有交互关系,所以我们认为两个候选蛋白质名称之间其他蛋白质名称的数量对交互作用关系的抽取会有一定的贡献。

(8) 两个候选蛋白质名称之间的距离特征。出现在两个候选蛋白质名称之间的词的个数,我们通过统计设定相应的阈值。直观上,两个候选蛋白质名称距离越近越有可能具有交互作用关系。

"We show here that recombinant bovine prion protein strongly interacts with the catalytic alpha/alpha' subunits of protein kinase."为 AImed 语料中的实例,我们从中提取的浅层语言学特征如表 2 所示。

表 2 浅层语言学特征实例

特征	特征描述	特征值
F_{p1}	第一个候选蛋白质名称词特征	p1 _ bovine, p1 _ prion, p1 _ protein
F_{p2}	第二个候选蛋白质名称词特征	p2 _ protein, p2 _ kinase
F_{pb}	两个候选蛋白质名称之间的词特征	b _ strongly, b _ interact, b _ with, …
F_l	左边三个词特征	l _ here, l _ that, l _ recombine
F_r	右边三个词特征	r _ .
F_k	关键词特征	interact _ between, distance _ 2
F_{ch}	两个候选蛋白质名称之间的语块中心词特征	chunk _ head _ strongly, chunk _ head _ interacts, chunk _ head _ with, chunk _ head _ ahpha/alpha', chunk _ head _ subunit, chunk _ head _ of
F_{ct}	两个候选蛋白质名称之间的语块类型	ADVP _ VP _ PP _ NP _ NP _ PP
F_{lc}	左边语块中心词特征	here _ that
F_{rc}	右边语块中心词特征	NULL
F_h	两个候选蛋白质名称的中心词对特征	prion _ kinase
F_s	两个候选蛋白质名称中心词的后缀对特征	ion _ ase
F_p	两个候选蛋白质名称中心词的前缀对特征	pri _ kin
F_{pn}	两个候选蛋白质名称间其他蛋白质的个数	0
F_d	两个候选蛋白质名称之间的距离	8,小于阈值,特征值设为 1

3 实验结果与分析

通常,研究人员把蛋白质交互作用关系抽取转化为二元分类问题。对句子中的任意两个蛋白质名称 Prot1 和 Prot2 所构成的实体对(Prot1, Prot2),使用分类器来判断是否构成预先定义的交互作用关系。本文采用由生物学家手工标注了人类蛋白质之间交互作用关系的 AImed 语料进行训练和测试,AImed 语料共包括 225 篇 Medline 摘要,其中 200 篇为正例即语料中包含蛋白质交互作用关系,其他 25 篇为负例。我们从包含两个及两个以上蛋白质名称的句子中产生实例,实例的数目由句子中蛋白质名称的个数 N 决定,每次从所有的蛋白质名称中选择两个,共产生实例 C_N^2 个。例如一个包含 3 个蛋白质名称的句子,可产生实例 $C_3^2 = 3$ 个。实验中共从 AImed 语料中产生实例 5070 个,其中包含蛋白质交互作用关系的正例有 1067 个。本文使用 SVM^{light} 工具集[14]进行了 3 组实验,实验结果的评价标准为精

确率(P),召回率(R)和 F 测度(F)。

3.1 实验 1

我们采用 SVM、TSVM 和 TSVM 与主动学习融合算法三种方法做对比实验,其中后两种方法把测试语料作为未标注语料加入到 TSVM 的训练过程中,采用 10 折交叉验证方法评测系统性能,实验结果如表 3 所示。表中上面 3 行是我们的实验结果,下面 6 行是当前在 AImed 语料上相关研究取得的较好结果,这些研究的实验设置和实验 1 完全相同,均是在 AImed 语料上采用 10 折交叉验证的方法评测系统性能。

表 3 AImed 语料实验结果

方法	P(%)	R(%)	F(%)
SVM	70.96	48.47	57.60
TSVM	58.19	65.89	60.71
TSVM + 主动学习	57.77	68.65	62.05
SVM-edit (Erkan et al., 2007)	77.52	43.51	55.61
SVM-cos (Erkan et al., 2007)	61.99	54.99	58.09
TSVM-edit (Erkan et al.,2007)	59.59	60.68	59.96
TSVM-cos (Erkan et al., 2007)	58.37	61.19	59.62
Rules (Yakushiji et al., 2005)	33.70	33.10	33.40
SVM (Mitsumori et al., 2006)	54.20	42.60	47.70

从实验结果看来,有监督的 SVM 性能与 Erkan 等[15]不相上下,但明显优于 Yakushiji 等[16]和 Mitsumori 等[17]的结果,表明本文选取的浅层语言学特征较好地体现了蛋白质交互作用关系的特点;TSVM 性能较 SVM 有了明显提高,表明 TSVM 有效利用了未标注语料的分布信息从而快速提高分类能力;TSVM 与主动学习融合算法取得了最好的结果,F 测度达到 62.05%,比未采用主动学习策略的 TSVM 方法 F 测度提高了 1.34%,这说明我们的样本选取策略是行之有效的。

3.2 实验 2

实验 1 在加入未标注语料规模较小的情况下,TSVM 性能有了较大提高,那么 TSVM 是否会在已标注语料较少而未标注语料较多的真实场合更加有效呢? 我们从 AImed 语料的 5070 个关系实例中随机抽取 10、20、50、100、200、500、1000、2000 和 3000 个实例作为训练语料,剩余的实例作为未标注的测试语料。实验中对以上各个规模的实例都随机抽取了 5 次,然后对 5 次所得的结果取平均值,实验结果如图 1所示。

从 F 测度曲线可以看出,在已标注语料小于 200 时,加入主动学习的 TSVM 较单一的 TSVM 性能提升不大,有时反而降低,但随着已标注语料规模的增大,主动学习使 TSVM 的性能不断得到提升,这说明主动学习可以在大规模已标注语料上发挥作用。同时,我们发现 SVM 在训练语料充足的情况下具有很强的泛化能力,但是随着已标注语料规模的减小,SVM 与 TSVM 性能的差距越来越大,TSVM 的优势也越明显。因此,在已标注语料有限而未标注自由文本易得的实际应用中,可以期待 TSVM 算法能发挥大的作用。

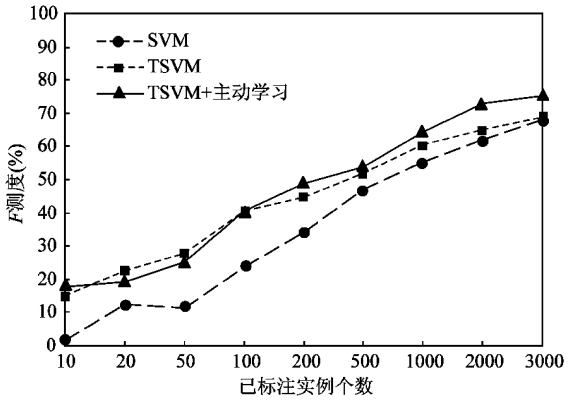

图 1 AImed 语料不同训练实例规模的 F 测度值

3.3 实验 3

在实验 2 中,我们认为 TSVM 可以在已标注语料较少而未标注语料较多的实际应用中更加有效,那么这个结论是否正确呢? 我们引入一个新的语料 BioCreAtIvE-PPI 语料①,该语料标注了 1000 个可能存在基因或蛋白质交互作用关系的句子,其中 173 个句子至少存在一个交互作用关系,589 个句子至少包含一个基因或蛋白质实体名称,语料总共标注了 255 个交互作用关系。

我们采取与 AImed 语料相同的方法从 BioCreAtIvE-PPI 语料中产生 1873 个关系实例,将这些实例作为未标注语料加入到 AImed 语料的 TSVM 训练过程中,10 折交叉验证的 F 测度为 54.76%。通过对比发现,这个结果比不加入 BioCreAtIvE-PPI 语料的 F 测度 60.71% 反而下降了约 6 个百分点。究其原因,我们认为 BioCreAtIvE-PPI 语料与 AImed 语料的分布不同,BioCreAtIvE-PPI 语料的分布信息在 TSVM 训练过程中转移到最终的分类器中,给分类结果带来了不利影响。

① http://www2.informatik.hu-berlin.de/~hakenber/corpora/#bc

为了解决这个问题,本文采用主动学习策略对自由文本语料进行有指导的优化选择。我们从 BioCreAtIvE-PPI 语料中分别随机抽取 0(即未加入新语料实例)、100、500、1000、1500 和 1873(即加入新语料全部实例)个实例连同测试语料一起作为未标注语料进行 TSVM 与主动学习融合实验,10 折交叉验证的 F 测度值如表 4 所示。

表 4　AImed 语料中加入不同规模新语料的 F 测度值

个数	0	100	500	1000	1500	1873
F(%)	62.05	62.24	62.77	63.58	63.91	64.12

从表 4 中可以看到,随着抽取的未标注实例个数的增加,*F* 测度值呈上升的趋势。当加入 BioCreAtIvE-PPI 语料中的所有实例时,*F* 测度达到 64.12%,与未加入新语料实例相比提高了 2.07 个百分点,同时比未采用主动学习策略(*F* 测度为 54.76%)高出 9.36 个百分点。实验结果表明,主动学习策略有效地选取了具有高训练效用的未标注样本,明显提高了系统的性能,较好地解决了 TSVM 中引入未标注自由文本进行有效学习的问题。

目前,本文提出的 TSVM 与主动学习融合的策略在生物医学文本中命名实体关系抽取上进行了实验并取得了较好的结果。实际上,该方法对于识别其他类型文本中的命名实体关系也具有普遍意义。例如,对于新闻等领域的文本也可以采取该方法来进行命名实体关系的识别[18]。同时,对于可以转化为分类问题的文本信息处理的一系列问题如语义角色标注[12]、语块识别[19]等,该方法均适用。

4　结　论

蛋白质交互作用关系抽取是生物医学文本信息抽取的重要组成部分,有着非常重要的应用价值,例如生物医学领域的问答系统,自动数据库生成系统,智能文档搜索和信息检索系统等。本文针对 TSVM 的特点,提出基于 TSVM 与主动学习融合的蛋白质交互作用关系抽取方法,通过有效的主动学习策略分别从候选未标注样本集和测试样本集中选取最优样本训练分类器。实验结果表明,本文的方法在小规模已标注语料环境下比有监督学习方法更优越;同时,通过加入不同分布的自由文本语料,验证了主动学习策略的有效性。由于实验中加入的自由文本规模不大,还不能完全体现 TSVM 与主动学习融合方法的优越性。我们认为,如果加入大量从不同环境中收集来的未标注自由文本,TSVM 与主动学习融合算法必然会取得更好的结果,这为今后的实际应用提供了一个较好的解决方案。

参考文献

[1] Xiao J, Su J, Zhou G D. Protein-protein interaction extraction: a supervised learning approach. In: Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine, Hinxton, Cambridgeshire, UK, 2005. 148-156

[2] Huang M, Zhu X, Hao Y, et al. Discovering patterns to extract protein-protein interactions from full biomedical texts. *BMC Bioinformatics*, 2004, 20(18): 3604-3612

[3] Donaldson I, Martin J, de Bruijn B, et al. PreBIND and textomy——mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 2003, 4: 11

[4] Haddow B, Matthews M. The extraction of enriched protein-protein interactions from biomedical text. In: Proceedings of Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics 2007, Prague, Czech Republic, 2007. 145-152

[5] Joachims T. Transductive inference for text classication using support vector machines. In: Proceedings of the 16th International conference on Machine Learning (ICML-99), Bled, Slovenia, 1999. 200-209

[6] Teng G F, Liu Y H, Ma J B, et al. Improved algorithm for text classification based on TSVM. In: Proceedings of the 1st International Conference on Innovative Computing, Information and Control (ICICIC-06), Beijing, China, 2006. 55-58

[7] Bruzzone L, Chi M, Marconcini M. A novel transductive SVM for the semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2006, 44(11): 3363-3373

[8] Kasabov N, Pang S N. Transductive support vector machines and applications in bioinformatics for promoter recognition. *Neural Information Processing-Letters and Reviews*, 2004, 2(3): 31-38

[9] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习.软件学报,2003, 14(3): 451-460

[10] Engelbrecht A P, Cloete I. Incremental learning using sensitivity analysis. In: Proceedings of IEEE International Joint Conference on Neural Networks, Washington DC, USA, 1999. 2. 1350-1355

[11] 张健沛,徐华.支持向量机(SVM)主动学习方法研究与应用.计算机应用,2004, 24(1): 1-3

[12] 陈耀东,王挺,陈火旺.半监督学习和主动学习相结合的浅层语义分析.中文信息学报,2008, 22(2): 70-75

[13] Tong S, Koller D. Support vector machine active learning with applications to text classification. In: Proceeding of the 17th International Conference on Machine Learning, Stanford, CA, USA, 2000. 401-412

[14] Joachims T. SVMlight support vector machine. http://svmlight.joachims.org: 2002

[15] Erkan G, Ozgur A, Radev D. Semi-supervised classification for extracting protein interaction sentences using dependency parsing. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Lzech Republic, 2007. 228-237

[16] Yakushiji A, Miyao Y, Tateisi Y, et al. Biomedical information extraction with predicate argument structure patterns. In: Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine, Hinxton, Cambridgeshire, UK, 2005. 93-96

[17] Mitsumori T, Murata M, Fukuda Y, et al. Extracting protein-protein interaction information from biomedical text with SVM. *IEICE Transactions on Information and Systems*, 2006, E89-D(8): 2464-2466

[18] 车万翔,刘挺,李生.实体关系的自动抽取. 中文信息学报,2005, 19(2): 1-6

[19] 梁颖红.基于多 Agent 的英汉文本语块识别技术研究:[博士学位论文]. 哈尔滨:哈尔滨工业大学计算机科学与技术学院,2006

Protein-protein interaction extraction based on combining TSVM and active learning

Liu Jianmiao[*,**], Wang Haochang[*,***], Zhao Tiejun[*]

([*] MOE-MS Key Laboratory of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin 150001)

([**] Department of Information Engineering, China Information Technology Designing & Consulting Institute Co., Ltd., Zhengzhou 450007)

([***] College of Computer and Information Technology, Daqing Petroleum Institute, Daqing 163318)

Abstract

This paper presents an algorithm for extraction of protein-protein interaction (PPI) based on the combination of the transductive support vector machine (TSVM) approach with the active learning algorithm to solve the problems which are the lack of labeled corpora and the easy usage of the vast amount of unlabeled biomedical free texts. The algorithm can maximally increase the performance of the TSVM algorithm through actively selecting useful unlabeled samples and adding them to the TSVM training set. The experiment results show that combing TSVM with the active learning is very promising on a mixed training set with a small number of labeled samples and a large number of unlabeled samples. Compared with the traditional support vector machine (SVM) algorithm and the TSVM algorithm, the paper proposed algorithm can immensely reduce the number of the training data and efficiently improve the performance of the classifier for PPI extraction. A very encouraging result of 64.12% F-score on the AImed corpus was achieved.

Key words: protein-protein interaction extraction, semi-supervised learning, transductive support vector machine (TSVM), active learning