文章编号:1000-5862(2013)04-0392-05

基于监督学习和半监督学习的蛋白质关系抽取

王艳华[1],杨志豪[1*],李彦鹏[1],唐利娟[2],林鸿飞[1]
(1. 大连理工大学计算机科学与技术学院,辽宁 大连　116024;2. 山东省农业管理干部学院机械电子工程系,山东 济南　250100)

摘要:提出了一种将监督学习和半监督学习融合的方法,并用于从文献中自动抽取蛋白质关系. 在 AImed 语料上的实验得到 63.2% 的 F 值,这表明该方法达到目前较好的性能.

关键词:文本挖掘;信息抽取;蛋白质关系抽取;监督学习;半监督学习
中图分类号:TP 391　　**文献标志码**:A

0　引言

过去的 10 年见证了生物医学文献的爆炸式增长,如生物文献的数据库 PUBMED 中的文献已经超过 2000 万条. 海量的数据给研究者们带来丰富的信息,但研究者们通宵达旦阅读文献也不及文献的增长速度. 因此,从生物医学文献中提取和组织信息的系统变得越来越重要. 抽取的这些信息能帮助研究者处理信息,系统地阐述生物模型、提出假设等. 随着研究的发展,从生物医学文献中自动抽取蛋白质间的相互作用关系,已成为文本挖掘领域中的重要方向.

使用机器学习方法抽取蛋白质间的相互作用关系是当前研究的热点,其中基于核的方法是一种特征抽取的有效方法,它保持对象的原始表达形式,通过计算一对实体的核函数的值使用这些对象. 许多核方法包括子序列核[1]、树核[2-3]、最短路径核[4]以及图核[5]已被用于蛋白质关系信息抽取. 然而,监督学习需要大量标注数据作训练集以保证泛化能力,而构建这些标注数据是非常费时费力的,特别是生物医学领域的标注通常需要标注者有一定的生物医学背景.

近年来提出的半监督学习是一种新的学习方法,它主要研究如何将大量的无标注样本和少量的有标注样本结合起来以提高学习器的泛化能力. 在生物信息学的实际应用中,对数据进行人工标注的代价很高,容易获取的是大量的廉价的未标注数据,如生物文献的数据库 PUBMED. 随着研究的发展,将少量带标注数据和大量的未标注数据结合的半监督学习成为机器学习的研究热点,大量的半监督学习方法纷纷涌现,如 M. Miwa 等[6]在 SVM 基础上提出的 TSVM (Transductive SVM) 和李彦鹏等[7]提出特征耦合泛化(FCG)方法,其优势是可以很容易处理大规模未标注数据,容易理解,易于实现,不受具体分类器的限制等. FCG 方法已被应用于生物医学的蛋白质关系抽取,并取得了较好的效果.

本文提出了一种基于将监督学习和半监督学习相融合的方法. 该方法融合了基于词特征的核、树核及图核 3 种监督学习方法和 FCG 半监督学习方法. 基于核函数(监督学习方法)的蛋白质关系抽取可以捕获结构化信息,取得了较高的性能,而 FCG(半监督学习方法)可以较好地解决数据稀疏问题. 鉴于 2 种方法的各自优势,本文将上述 3 种监督学习方法和 FCG 半监督学习方法融合,在 AImed 语料上获得了较好的性能.

1　研究方法

监督学习方法利用样本的标注信息能达到较好的分类性能,但却忽视了大量的未标注数据蕴含的丰富的信息. 因此,将监督学习方法和半监督学习方法结合起来,即利用了标注信息又考虑到了未标注

收稿日期:2012-11-15
基金项目:国家自然科学基金(61070098,61272373)和中央高校基本科研业务费专项资金(DUT13JB09)资助项目.
通信作者:杨志豪(1973-),男,黑龙江大庆人,副教授,博士,博士生导师,主要从事文本挖掘领域的研究.

数据,以此来提高蛋白质关系抽取的性能. 使用的监督学习的方法有 3 种: 基于特征的核、图核、树核,使用的半监督学习方法是特征耦合泛化. 最后对多个方法的标准化结果求和,来将监督学习方法和半监督学习方法融合.

1.1　基于词特征的核

本文使用的词特征包括以下 4 种:

(ⅰ) 蛋白质实体的区域特征: 该特征不但标明某个 N 元组是否出现在句子中而且还指明了该 N 元组出现在第 1 个蛋白质左边,2 个蛋白质之间还是第 2 个蛋白质右边. 如特征"Word_Left = expression"表示"expression"这个词出现在第 1 个蛋白质的左边. 本实验中 N 设为 1 ~ 3.

(ⅱ) 蛋白质实体的距离特征: 该特征赋予在蛋白质周围的每个 N 元组具体的位置信息. 如特征"-1_From_P1 = expression"表示第 1 个蛋白质左边的第 1 个词是"expression". 本实验中 N 设为 1 ~ 3.

(ⅲ) 交互词特征: 在大多数情况下,如果 2 个蛋白质实体之间存在某种关系,那么这 2 个蛋白质实体周围会出现某些动词或者它们的变体,如"bind"、"interaction"、"inhibit". 这些词的出现往往以较大的概率值判断出这 2 个蛋白质实体之间存在交互关系. 因此引入布尔特征,1 表示该特征出现,0 表示该特征不出现.

(ⅳ) 否定词特征: 在交互词特征出现的情况下,它的前面时常伴随"not","neither","no"等否定词. 如果出现交互词就认定 2 个蛋白质实体有交互关系,可能会造成很高的错误率,因此引入了否定词特征. 该特征也是布尔特征,在交互词出现时,若否定词出现特征值为 0,否则为 1.

1.2　图核

图核的主要目的是根据句法分析树的树状及结构分析出 2 个蛋白质的距离信息,进而将句子表示为 1 个图结构,并建立矩阵计算相似度. 使用的图核方法是由 A. Airola 等提出的全路径图核. 图核中主要使用了 2 类子图: 分析结构子图(PSS)和线性顺序子图(LOS). 图 1 列举了 1 个图的表达实例,图 1 的上半部分是 PSS,下半部分是 LOS.

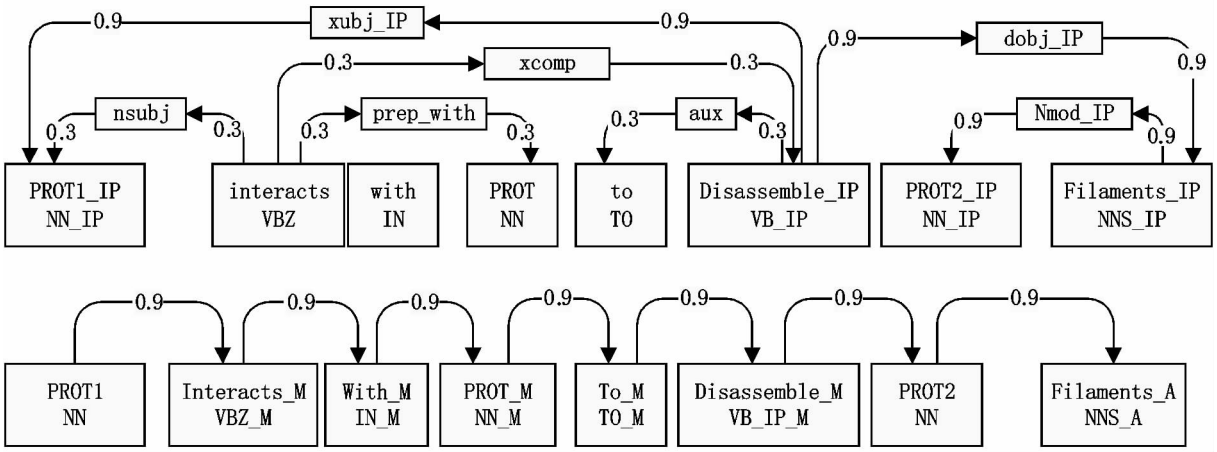

图 1　图核的 1 个表达实例

图核方法通过比较 2 个图的顶点标签关系以计算句子间的相似度,具体可以用图矩阵 $\boldsymbol{G}$ 表示为

$$\boldsymbol{G} = \boldsymbol{L}^{\mathrm{T}} \sum_{n=1}^{\infty} A^n \boldsymbol{L}, \tag{1}$$

其中 $\boldsymbol{L}$ 是 1 个 $N \times L$ 的标签矩阵,矩阵 $\boldsymbol{A}$ 是 1 个 $N \times N$ 的边矩阵,N 表示顶点数目,L 表示标签数目. L_{ij} 为 1 表示顶点 i 含有标签 j,否则为 0; A_{ij} 中的每个元素 a_{ij} 标定权重,如果顶点 i 和顶点 j 是连通的就赋予权重 0.9 或 0.3,否则标注 0.

公式(1) 计算了任何一对蛋白质关系对各个顶点之间的所有路径的权重和,所以,每条记录都表示了一对顶点之间的关系强度. 使用 2 个图矩阵 $\boldsymbol{G}$ 和 $\boldsymbol{G}'$ 作为输入,图核 $k(\boldsymbol{G},\boldsymbol{G}')$ 的计算公式为

$$k(\boldsymbol{G},\boldsymbol{G}') = \sum_{i=1}^{L} \sum_{j=1}^{L} G_{ij} G'_{ij}. \tag{2}$$

1.3　树核

树核是通过计算 2 个句法分析树 T_1 和 T_2 的相同子树结构的数目作为二者的语义相似度的一种方法. 卷积树核是一种典型的树核,卷积树核 $K_C(T_1,$

T_2) 是通过计算公共子树的数量作为句法结构的相似度[12]

$$K_C(T_1,T_2) = \sum_{n_1\in N_1,n_2\in N_2}\Delta(n_1,n_2) \quad (3)$$

其中 N_j 是树 T_j 中的节点集,$\Delta(n_1,n_2)$ 使用以下递归的算法计算以 n_1 和 n_2 为根的相同的子树结构数目:

(ⅰ) 如果 n_1 和 n_2 不同即没有上下文信息,则使用上下文无关文法(CFG) 规则,令 $\Delta(n_1,n_2)=0$;否则执行(ⅱ).

(ⅱ) 如果 n_1 和 n_2 都是 POS 标签,则 $\Delta(n_1,n_2)=1\times\lambda$;否则执行(ⅲ).

(ⅲ) 使用递归计算 $\Delta(n_1,n_2)$:

$$\Delta(n_1,n_2) = \lambda\prod_{k=1}^{\#ch(n_1)}(1+\Delta(ch(n_1,k),ch(n_2,k))), \quad (4)$$

其中 $\#ch(n)$ 是节点 n 的孩子的个数,$ch(n,k)$ 是节点 n 的第 k 个孩子,$\lambda(0<\lambda<1)$ 为衰减因子,使对不同大小的子树下的核函数值稳定.

1.4 特征耦合泛化(FCG)

特征耦合泛化是一种半监督学习的方法. FCG 方法的核心思想是:利用实例区分特征(EDF) 和类别区分特征(CDF) 在未标注语料中的共现来产生新的特征,其算法如图 2 所示.

输入:原始数据集合 $D=\{d_1,d_2,\cdots,d_m\}$,原始特征集合 $F=\{f_1,f_2,\cdots,f_n\}$,向量集表示 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$,未标注数据集 U.

(1) 从原始特征集 F 中选择子集 E 作为 EDF 集合.

(2) 确立映射函数 $root(e)=E\to R$,将 E 中的元素映射到高级概念集合 R.

(3) 从原始特征集合 F 中选择子集 C 作为 CDF 集合.

(4) 定义 FCD 类型集合 T 来度量每个 EDF 和 CDF 的耦合程度.

(5) 根据未标注数据 U 和 FCD 类型 T,计算得到每个 EDF 和 CDF 之间的 FCD 值.

(6) 将集合 $G=R\times C\times T$ 作为新特征集合,使得每个新特征对应一个 3 元组 (r,c,t),其中 $r\in R,c\in C,t\in T$.

(7) 将每个实例 $d\in D$ 从原始特征向量表示 $\boldsymbol{x}\in X$ 转化成新的特征向量 $\tilde{\boldsymbol{x}}$,其中 $\tilde{x}_i\in\tilde{\boldsymbol{x}}$ 的值按如下公式计算:

$$\tilde{x}_i=\tilde{x}_{(r,c,t)}=\sum_{e\in E\&root(e)=r}BFeature(e,d)\times FCD(U,e,c),$$

其中 i 是新特征的索引,与 G 中每个 3 元组 (r,c,t) 一一对应. 当样本 d 满足特征 e 时函数 $BFeature(e,d)$ 的值为 1,否则为 0.

输出:新的特征集合 G,转化后的特征向量 $\tilde{\boldsymbol{x}}=\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_m\}\subset R^{dim(R)*dim(C)*dim(T)}$

图 2 特征耦合泛化算法

使用特征耦合算法生成新的特征向量后,使用支持向量机(SVM) 进行分类. 在 SVM 的使用中,用到了 2 种核. 一种是线性核(在实验部分记作 FCG-线性核);另一种是 RBF 核,在此之前对得到的特征向量使用 SVD(Singular Value Decomposition,奇异值分解) 进行了降维,这种方法在实验部分记作 FCG-SVD-RBF. 这种子空间方法和非线性分类其方法的组合是模式识别、图像识别里经典的分类方法[8-9],在很多任务中取得了比线性方法更好的效果,子空间降维的方法可以在一定程度上减少模型复杂度,而非线性分类器通过提升复杂度来减少训练错误率,二者相互制约相互促进从而达到更好的学习和泛化效果.

1.5 监督学习方法和半监督学习方法的融合

每种方法都有着各自的优点或缺点. FCG 能很好的解决数据稀疏问题,基于特征的核(在实验中使用 SVM 分类器的线性核) 简单有效,但两者都不能捕获句子的结构信息. 图核、树核能较好利用句子结构信息,从不同的角度计算句子的相似度,融合这些相似度可以避免遗漏重要的特征. 为了实现基于监督学习和半监督学习方法的融合,对多个方法 K_m 的标准化结果求和,其中 m 代表了方法的个数:

$$K(x,x') = \sum_{m=1}^{M}\sigma_m K_m(x,x'),$$

$$\sum_{m=1}^{M}\sigma_m = 1, \sigma_m \geqslant 0. \quad (5)$$

2　实验

2.1　实验设置

在本文的实验中，未标注语料是 1979—2009 年的 PUBMED 文摘，标注语料是公共评测语料 AImed. 该语料具有较大的规模，适合用于机器学习方法的训练，近年来已成为蛋白质关系抽取方法的评测标准. 监督学习和半监督学习均用 10 倍交叉检验的方法进行实验. 在 FCG-SVD-RBF 的实验中，当惩罚因子 c 设置为 3，RBF 核的 γ 参数设置为 26 时实验效果最佳. 当词特征参数惩罚因子 $c = 0.03$ 时实验效果最好. 其它方法中 SVM 设置为默认参数.

2.2　实验结果和讨论

实验采用的性能评测指标是当前 PPI 抽取系统主要使用的准确率（P）、召回率（R）、F 值（F），$F = (2PR)/(P+R)$. 分别计算 10 倍交叉验证每次的 P、R、F 值，最后将 10 次结果的宏平均值最为最终的 P、R、F 值. 监督学习方法，半监督学习方法及 2 种学习方法结合的性能如表 1 所示. 其中，括号中的数字代表该方法在融合时的权重，通过反复的实验选择的权重能使融合的效果最佳.

半监督学习方法包括 FCG-线性核和 FCG-SVD-RBF（将 2 种方法结合起来统称为 FCG-all）. 监督学习方法包括基于词特征的核、图核和树核. 如表 1 所示，可以看到半监督学习方法中先用 SVD 降维，然后再用 SVM 的 RBF 核的实验结果要优于 FCG-线性核. 而两者的结合取得了更好的结果（59.2% 的 F 值）. 在监督学习方法中，基于词特征的核的性能最好. 将半监督学习方法和监督学习方法融合起来的 F 值为 63.2%，优于单独使用一种或几种监督学习方法（或半监督学习方法）的结果.

表 1　监督学习、半监督学习方法及 2 种学习方法结合的性能

方法	P	R	F(%)
FCG-线性核	56.6	58.8	56.8
FCG-SVD-RBF	55.6	63.6	58.2
FCG-all	58.1	61.8	59.2
基于词特征的核	57.1	67.2	61.2
图核	49.1	67.4	56.3
树核	45.0	57.4	48.3
词特征(0.75) + 图核(0.05) + 树核(0.2)	57.8	69.1	62.0
FCG-all(0.2) + 词特征(0.65) + 图核(0.05) + 树核(0.1)	59.6	68.0	63.2

表 2 是本系统与当前性能最好的几种 PPI 抽取方法之间的性能对比. A. Airola 等使用全路径图核进行蛋白质关系抽取；杨志豪等[10]的方法融合了基于特征的核、树核和图核，并扩展了最短路径依存树以及依存路径；M. Miwa 等[11]融合了多层语义信息，通过基于多种语法分析器的多核的合并实现蛋白质关系抽取.

表 2　不同方法在 AImed 语料上的性能

方法	P	R	F(%)
本文的方法	60.0	67.9	63.2
文献[10]的方法	57.1	70.3	63.9
文献[11]的方法	60.4	69.3	63.5
文献[5]的方法	52.9	61.8	56.4

与这些当前最优性能的系统比较，也获得了与其接近的性能. 原因在于监督学习方法利用标注数据训练模型能取得较好的效果，而半监督学习方法能利用大量未标注数据中的信息，将监督学习方法和半监督学习方法融合起来，充分发挥两者的优势，能取得比两者单独使用更好的效果.

3　结论

本文提出一种用于从生物医学文献中自动抽取蛋白质关系的方法. 这种方法将基于词特征的核、树核、图核的 3 种监督学习方法及半监督学习方法特征耦合泛化进行相应的融合，在 AImed 语料上的实验得到 63.2% 的 F 值. 这表明将半监督学习和监督学习融合可以达到很好的互补效果.

4　参考文献

[1] Bunescu R C, Mooney R J. Subsequence kernels for relation extraction [C]. MA: MIT Press, 2006: 171-178.

[2] Moschitti A. Making tree kernels practical for natural language processing [C]. NJ: Association for Computational Linguistics, 2006: 113-120.

[3] Sætre R, Sagae K, Tsujii J. Syntactic features for protein-protein interaction extraction [C]. Germany: CEUR, 2007.

[4] Bunescu R C, Mooney R J. A shortest path dependency kernel for relation extraction [C]. NJ: Association for Computational Linguistics, 2005: 724-731.

[5] Airola A, Pyysalo S, Bj? rne J, et al. All-paths graph kernel for proteinrotein interaction extraction with evaluation of cross corpus learning [EB/OL]. [2012-10-13]. http://www.biomedcentral.com/1471-2105/9/S11/S2.

[6] Miwa M, Sætre R, Miyao Y, et al. A rich feature vector for protein-protein interaction extraction from multiple corpora [EB/OL] [2012-10-15]. http://www.aclweb.org/anthology-new/D/D09/D09-1013.pdf.

[7] Li Yanpeng, Hu Xiaohua, Lin Hongfei, et al. Learning an enriched representation from unlabeled data for protein-protein interaction extraction [EB/OL]. [2012-10-17]. http://www.biomedcentral.com/1471-2105/11/S2/S7.

[8] Duda RO, Hart PE. Pattern classification and scene analysis [M]. New York: John Wiley, 1973.

[9] 刘立月,黄兆华,刘遵雄. 高维数据分类中的特征降维研究 [J]. 江西师范大学学报: 自然科学版, 2012, 36(2): 131-134.

[10] 唐楠,杨志豪,林鸿飞,等. 基于多核学习的医学文献蛋白质关系抽取 [J]. 计算机工程, 2011, 37(10): 184-186.

Protein-Protein Interaction Extraction Based on the Combination of Supervised and Semi-Supervised Learning Method

WANG Yan-hua[1], YANG Zhi-hao[1*], LI Yan-peng[1], TANG Li-juan[2], LIN Hong-fei[1]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian Liaoning 116024, China;
2. Department of Mechanical and Electronic Engineering, Shandong Agricultural Administrators College, Ji´nan Shandong 250100, China)

Abstract: An approach based on supervised learning and semi-supervised learning to automatically extract protein-protein interactions from biomedical literature has been presented. Experimental evaluations show that the method can achieve the state of art performance with 63.2% *F*-score on the AImed corpus.

Key words: text mining; information extraction; protein-protein interaction; supervised learning; semi-supervised learning

(责任编辑: 冉小晓)