

文章编号: 1003-0077(2016)01-0024-06

基于多核融合的中文领域实体关系抽取

郭剑毅^{1,2}, 陈鹏¹, 余正涛^{1,2}, 线岩团^{1,2}, 毛存礼^{1,2}, 赵君¹

(1. 昆明理工大学 信息工程与自动化学院, 云南 昆明 650500;
2. 昆明理工大学 智能信息处理重点实验室, 云南 昆明 650500)

摘要: 针对传统径向基核函数的训练矩阵中所有元素都十分接近零而不利于分类的问题, 该文提出了一种融合了改进的径向基核函数及其他核函数的多核融合中文领域实体关系抽取方法。利用径向基核函数的数学特性, 提出一种改进的训练矩阵, 使训练矩阵中的向量离散化, 并以此改进的径向基核函数融合多项式核函数及卷积树核函数, 通过枚举的方式寻找最优的复合核函数参数, 并以上述多核融合方法与支持向量机结合进行中文领域实体关系抽取。在旅游领域的语料上测试, 相对于单一核方法及传统多核融合方法, 关系抽取性能得到提高。
关键词: 关系抽取; 径向基核函数; 卷积核函数; 多核融合
中图分类号: TP391 **文献标识码:** A

Domain Specific Chinese Semantic Relation Extraction Based on Composite Kernel

GUO Jianyi^{1,2}, CHEN Peng¹, YU Zhengtao^{1,2}, XIAN Yantuan^{1,2}, MAO Cunli^{1,2}, ZHAO Jun¹
(1. The School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China;
2. Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: This paper proposes a composite kernel approach to Chinese semantic relation extraction by a composite kernel. This paper designs an improved training matrix by using the mathematical properties of radial basis kernel in order to make vectors disperse in the training matrix, then integrate this kernel with the polynomial kernel and the convolution tree kernel. It enumerates for the best parameters of the composite kernel function for Chinese semantic relation extraction. Experimented on the tourist domain texts, the proposed method out-performs methods of single kernel as well as a traditional composite kernel.
Key words: relation extraction; radial basis kernel function; convolution kernel; composite kernel

1 前言

中文领域实体关系抽取即从多样化的中文领域文本中找出实体对之间的关系, 是自然语言处理的基础, 为中文领域信息检索、自动问答系统、机器翻译、本体构建等提供重要技术支持。
目前中文领域关系抽取的基于机器学习方法主要包括基于特征向量^[1-2]的方法、基于卷积核函数^[3-5]及多核融合^[6-8]的方法。基于特征向量的机器学习方法通过构造分类器来进行关系分类, 虽然运

算速度快, 但是仅利用平面核函数表达文本的平面信息因为不能有效挖掘文本的结构信息, 使抽取性能遇到了瓶颈。利用卷积核函数替代平面核函数计算两个对象的相似度, 在充分挖掘句法信息或依存信息等文本结构信息方面具有一定的潜力, 因此卷积核函数方法应用在中文领域实体关系抽取上取得了一定的发展。Yu 等^[3]为了增加句法树包含的信息, 提出基于卷积树核函数的中文实体语义关系抽取方法, 构造能有效捕获结构化信息和实体语义信息的合一句法和实体语义关系树, 从而提高了中文语义关系抽取的性能; Liu 等^[5]在原有关系实例的

最短路径包含树的基础上,利用 Hownet 加入语义信息,从而提高中文关系抽取性能。但是仅使用单一的卷积核方法忽略了文本的平面信息具有局限性。多核融合的方法兼顾了文本的平面信息及结构信息的优点,目前在关系抽取领域取得了很好的效果;Huang 等^[6]提出了一种卷积树核分别与线性核及多项式核融合的多核融合方法,表明多核融合能有效提高关系抽取的性能;文献^[7]证明,在多核融合的过程中,提高单一核函数在关系抽取中的性能能够提高复合核函数在关系抽取中的性能。由于传统的径向基核函数在形成训练矩阵时,训练矩阵中的所有元素趋近于 0 而导致分类模型分类效果不佳,使得融合了径向基核函数的复合核函数的中文领域关系抽取性能下降。

针对上述问题,本文以中文旅游领域为对象,根据径向基函数的数学特征,改造径向基核函数训练矩阵,以解决传统径向基核函数训练矩阵的元素趋近于 0 而不利于分类的问题,融合了多项式核函数及卷积树核函数,试图提高中文领域关系抽取的性能。实验结果表明,在中文旅游领域语料下,采用本文提出的模型进行关系抽取,其准确率、召回率和 F 值均优于单一核函数方法及单一平面核函数与树核函数融合的方法。

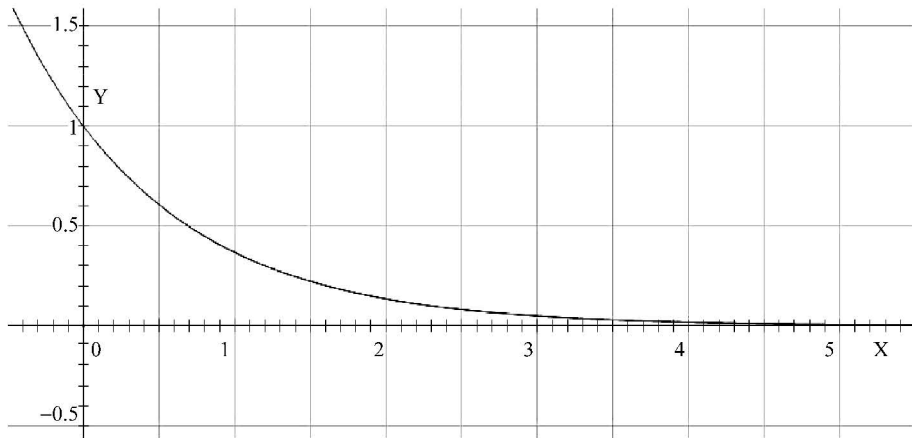


图 1 径向基函数曲线图

对于任意两个特征向量 $X_i, X_j, i, j \in \{1, 2, \dots, m\}$, 它们的映射构成新的映射空间中第 i 个向量的第 j 个特征, 而由于径向基核函数的数学性质, 使得每一个元素都十分趋近于 0, 这是十分不利于进行分类工作的。例如, 有三个二维向量 a, b, c 分别为: $a = (2, 3), b = (4, 10), c = (3, 5)$, 它们映射到径向基核函数对应的空间的训练矩阵为:

2 基于多核融合的领域实体关系抽取

2.1 改进的复合平面核函数

2.1.1 核函数训练矩阵

对于平面特征向量形成的特征矩阵, 在被某种映射规则映射后形成的矩阵即为核函数训练矩阵, 训练矩阵是一个半正定对称矩阵。假设有特征向量 $X_1, X_2, \dots, X_m, i, j \in \{1, 2, \dots, m\}$ 。有映射 $\phi(X)$, 且 $\phi(X)$ 的点积 $K(X_i, X_j)$ 为合法的核函数。则它们经过映射 ϕ 形成的训练矩阵:

$$K_{\text{training}} = \begin{bmatrix} k(X_1, X_1) & \dots & k(X_1, X_m) \\ \dots & \dots & \dots \\ k(X_m, X_1) & \dots & k(X_m, X_m) \end{bmatrix}$$

在之后的分类过程中, 核函数矩阵是唯一能被运用的信息。也就是说, 核函数训练矩阵质量的好坏决定了分类的好坏, 从而影响关系抽取的质量。

2.1.2 改进的径向基训练矩阵

径向基核函数是一种平移不变核函数, 具体表达式为式(1)所示。

$$Y(x) = \exp(-a \cdot x + b) \quad (1)$$

在系数 $a=1, b=0$ 下, 径向基函数如图 1 所示。由图 1 可以看出, 随着 x 增加, 函数值极其迅速的接近于 0。

$$K_{a,b,c} = \begin{bmatrix} 0 & 9.602 \times 10^{-24} & 6.738 \times 10^{-3} \\ 9.602 \times 10^{-24} & 0 & 5.109 \times 10^{-12} \\ 6.738 \times 10^{-3} & 5.109 \times 10^{-12} & 0 \end{bmatrix}$$

为了解决上述问题, 本文采用将训练矩阵中每一个特征都限制在一个适当的范围内, 而便于进行分类工作。例如, 将范围限制在 0.2~1 之间, 改进训练矩阵的方法为:

$$1) \text{ 计算 } Feature_{ij} = \|X_i - X_j\|^2, \text{ 其中 } i, j \in$$

$\{1,2,\cdots,m\}$;

2) 枚举所有的 Feature_{ij} , 找到最大值, 记为 Feature_{\max} ;

3) 计算常数 $\delta = (-\ln 0.2) / \text{Feature}_{\max}$;

4) 规范化后的训练矩阵即: $K_{\text{new}} = [k(\text{Feature}_{ij} \times \delta)]_{m \times m}$

其中 $i, j \in 1, 2, \cdots, m$ 。将上述三个向量 a, b, c 规范化处理后的训练矩阵为:

$$K_{\text{new}} = \begin{bmatrix} 0 & 0.200\ 0 & 0.859\ 3 \\ 0.200\ 0 & 0 & 0.454\ 1 \\ 0.859\ 3 & 0.454\ 1 & 0 \end{bmatrix}$$

2.1.3 基于径向基与多项式复合平面核函数

由于大部分平面核函数都可以由内积核及平移不变核表示, 而不同的核函数空间分类效果也不同。为了融合内积核函数及平面核函数的特性, 本文选择多项式核函数及改进的径向基核函数的线性复合平面核函数表达平面信息。定义如式(2)所示。

$$CPK(V_1, V_2) = \beta RBF(V_1, V_2) + (1 - \beta) PK(V_1, V_2) \quad (2)$$

其中 $CPK(V_1, V_2)$ 为复合平面核函数, $RBF(V_1, V_2)$ 为径向基核函数, $PK(V_1, V_2)$ 为多项式核函数, β 为复合平面核函数权重参数, V_i 为任意实例的特征向量。

本文特征集选择了中文领域实体关系抽取中成熟的特征^[1-2,6-7]。特征选择如下: 实体类型、实体对的组合类型、实体词性及上下文词性(窗口为 2)^[1]、实体对距离、实体对间出现其他实体的个数、实体对间语义词汇。结合以上特征选择, 将实例向量化, 形成特征矩阵。当获取了复合平面核函数矩阵集后, 将遍历训练其中所有复合平面核函数, 获取不同核函数对应的分类器, 分别考察分类器的抽取性能, 以确定最优复合平面核函数。

2.2 融合语义信息的卷积树构造

本文采用 Collins 和 Duffy^[9] 的卷积树核函数

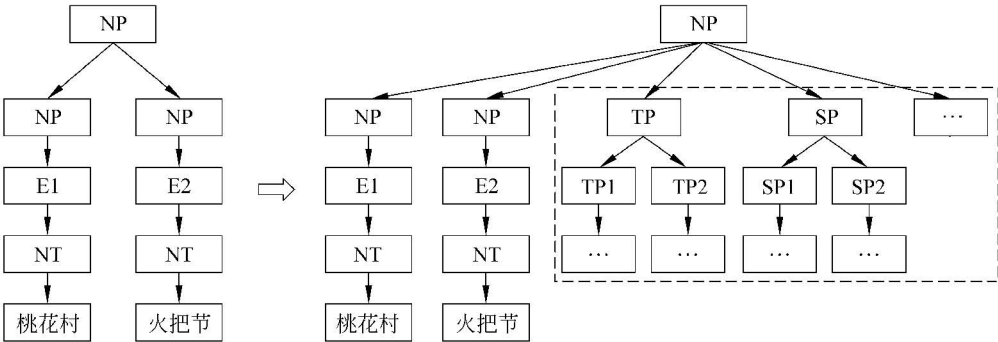


图 2 集成树的关系实例

(Convolution Tree Kernel,CTK),即两棵树之间的相似度可以通过计算它们之间的相同子树的数目来实现,表示为公式(3)。

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

其中 N_1 和 N_2 分别为 T_1 和 T_2 的节点集合, $\Delta(n_1, n_2)$ 用来计算以 n_1 和 n_2 为根节点的两颗子树之间的相似度。

2.2.1 基于句法树的实例表示方式

卷积树核函数计算的对象为两个子树,在实体语义关系抽取中,关系实例的表达方式有多种。一个完全句法树结构过于复杂,包含了大量的噪声信息,因此基于完全句法树的实例的关系抽取效果不理想,因此通常需要对句法树进行裁剪。Zhang^[10]等在效果最好的最小完全句法树(MCT)和最短路径树(PT)的实验表明,PT 树结构的实验结果最好。故本文采用 PT 树为基础融合语义信息。

2.2.2 基于句法树和语义树的集成树

在句法树结构中,并不包含实例的语义信息,如:实体大类类型、实体子类类型和实体引用类型等。文献[3,11]探讨了语义信息加入到句法树中对关系抽取性能的影响,加入语义信息后的树能够同时包含结构化及平面化的信息,从而有更好的抽取性能。所以,本文对句法树进行改造加入语义信息。根据已有的研究,本文以特征匹配树(FTP)为添加方式,将语义信息直接挂在根节点上,加入实体大类,实体小类及引用类型三种语义信息。

集成树构造的具体过程为:首先将同一个关系中的两个实体的同一特征先挂到父节点上,然后再统一挂到根节点上。例如,“TP1”和“TP2”先挂到父节点“TP”上,然后再同其他特征节点挂到根节点上,如图 2 所示。

2.3 基于平面核及卷积核的复合核函数

根据已定义的卷积树核 CTK 和复合平面核函数 CPK,定义复合核函数如式(4)所示。

$$CK(R_1,R_2) = \alpha CTK(T_1,T_2) + (1-\alpha)CPK(V_1,V_2)$$
 (4)

其中,CK(R₁,R₂)表示两个实例 R₁、R₂之间的相似度;CPK(V₁, V₂) 表示两个实例 R₁和 R₂的两个特征向量 V₁和 V₂的相似度,可由复合平面核计算得到;CTK(T₁, T₂)是由卷积树核计算得到两棵子树的相似度,T₁和 T₂是从两个候选关系实例的分析树中抽取出来的子树。@ 是复合核函数权重参数。

3 实验设计与结果分析

相对于其他机器学习算法,支持向量机在实体关系抽取领域通用性较好,分类精度高,且分类速度只与支持向量数目而不是所有的训练样本数目有关,所以本文选择支持向量机训练数据,本文的实验工具使用 Tree Toolkits^①。本文使用的语料为人工从互联网及文献资料中获取的中文旅游文本共 600 余篇,预处理后包括正例 1 023 个,负例 5 450 个。在训练中使用十倍交叉验证以最大化利用数据,实验评测采用自然语言处理的通常使用的标准:准确率、召回率、F 值, F 值评测系统的最终性能。

3.1 实验设计

为了验证本文方法的有效性,并与其他传统方法进行比较,本文设计了四项任务。

任务 1,验证改进的径向基核函数对关系抽取性能的影响。由于径向基函数的特性,在函数值为 0.1 后函数趋近为 0 的趋势很快,所以设定规范的范围分别为 0~1,0.1~1,0.2~1,0.3~1 四组范围;任务 2,验证复合平面核函数对抽取性能的影响与平面核树及树核形成的复合核对抽取性能的影响成正比。实验用枚举的方法探索权重参数 β 对复合平面核抽取性能的影响;任务 3,利用枚举的方法寻找树核函数及复合平面核函数在融合形成的复合核函数中占什么比例,使得抽取性能最优;任务 4,验证本文提出系统的有效性。用任务 3 找到的最优比例的多核融合的复合核核函数与特征向量方法,单一核方法及其他多核融合系统进行比较。

3.2 实验结果及分析

3.2.1 验证改进的径向基训练矩阵对抽取性能的影响

由表 1 可以看出,系统经过训练矩阵规范化处理后,准确率和召回率都有较大的提高,但是并不是随着取值范围的最小值增加系统的 F 值就增加。说明规范范围缩小到一定程度后,系统分类效果降低。在 0.2~1 这个范围内系统 F 值最高,本文以后的实验中,径向基核函数的训练矩阵都以这个范围规范。

表 1 不同训练矩阵下的 RBF 抽取性能

规范范围	评价指标/%		
	准确率	召回率	F 值
0-1	55.60	39.85	46.43
0.1-1	61.33	<u>53.25</u>	57.01
0.2-1	<u>62.50</u>	53.19	<u>57.47</u>
0.3-1	59.20	52.67	55.74

3.2.2 寻找复合平面核函数的最优比例

实验中设置权重参数 β 上下限为 0~1,步长为 0.1。特别的,在不考虑 α 的情况下,当 β 为 0 时,平面核函数只包括多项式函数的特性;当 β 为 1 时,平面核函数只包括径向基核函数的特性。由于只考虑权重参数 β 对抽取性能的影响,这里假设 α 为 0.5。

表 2 不同权重参数 β 下多核融合核函数的抽取性能

权重参数 β	评价指标/%		
	准确率	召回率	F 值
0	75.26	52.23	61.66
0.1	<u>77.36</u>	57.26	<u>65.81</u>
0.2	74.28	<u>57.33</u>	64.73
0.3	71.55	55.00	62.19
0.4	68.36	55.66	61.36
0.5	65.25	55.66	60.01
0.6	64.33	55.00	59.30
0.7	66.00	55.00	60.00
0.8	63.28	55.00	58.85
0.9	65.55	54.89	59.75
1.0	66.89	53.28	59.31

① <http://disi.unitn.it/moschitti/Tree-Kernel.htm>

从表 2 可以看出,对于旅游领域,当权重参数 β 为 0.1 时,性能是最好的,并且召回率对于权重参数 β 并不十分敏感。

3.2.3 寻找平面核与树核融合的最优比例

从表 3 可以看出,对于旅游领域,当权重参数 α 为 0.3 $\alpha_1 = 0.1$ 时,抽取性能是最好的,说明平面核函数对多核融合核函数抽取性能的贡献要更多一些。

表 3 不同权重参数 α 下多核融合核函数的抽取性能

权重参数 α	评价指标/%		
	准确率	召回率	F 值
0	74.29	59.32	65.97
0.1	75.23	57.49	65.17
0.2	75.50	53.66	62.73
0.3	<u>78.25</u>	<u>58.10</u>	<u>66.69</u>
0.4	76.56	53.02	62.65
0.5	77.36	57.26	65.81
0.6	75.56	50.47	60.52
0.7	74.60	48.28	58.62
0.8	76.68	48.55	59.46
0.9	74.25	49.60	59.47
1.0	73.56	32.83	45.74

3.2.4 总体性能与其他同类系统的比较

表 4 与其他同类系统的比较

方法	评价指标/%		
	准确率	召回率	F 值
单一核方法			
线性核函数	68.70	47.94	56.16
卷积树核	73.56	32.83	45.40
多核融合方法			
复合平面核	74.29	55.32	63.42
树核+多项式核	75.26	52.23	61.66
树核+径向基核	66.89	53.28	59.31
本文提出的方法	<u>78.25</u>	<u>58.10</u>	<u>66.69</u>

在表(4)这一组实验中可以看出:(1)无论是复合平面核函数还是树核函数,其性能都没有两种核函数融合的复合核函数抽取性能好;(2)与树核与单一核函数的复合核函数比较,当提高了平面核函数

性能后,整体关系抽取性能有所增加。

4 结束语

本文以支持向量机为机器学习算法,通过改进的径向基核与多项式核及卷积树核融合得到多核融合核函数,进行中文领域实体关系抽取。在中文旅游领域中,本文提出的多核融合系统取得了 66.69% 的 F 值,与单一核及其他平面核与树核的复合核函数相比,抽取性能有所提高。下一步工作中,将尝试以下两种途径以期待抽取性能的提高:

- (1) 融入多样化的核函数以提高抽取能力;
- (2) 提高单一核函数的抽取能力以提高整体抽取能力。

参考文献

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

[2] Lei Chunya, Guo Jianyi, Yu Zhentao, et al. The Field of Automatic Entity Relation Extraction based on Binary Classifier and Reasoning [C]//Proceedings of the Third International Symposium on Information Processing. Qingdao, China, 2010:327-2-331.

[3] 虞欢欢,钱龙华,周国栋,等. 基于合一句法和实体语义树的中文语义关系抽取[J],中文信息学报,2010,24(5):17-23.

[4] Peng Cheng, Gu Jinghang, Qian Longhua. Research on Tree Kernel-Based Personal Relation Extraction [J]. Communications in Computer and Information Science,2012, 333:225-236.

[5] Liu Dandan, Zhao Zhiwei, Hu yanan, et al. Incorporating Lexical Semantic Similarity to Tree Kernel-based Chinese Relatin Extraction[J]. Lecture Notes in Computer Science, 2013, 7717: 11-21.

[6] 黄瑞红,孙乐,冯元勇,等. 基于核方法的中文实体关系抽取研究[J]. 中文信息学报,2008,22(5):102-108.


[7] Zhang Ji, Ouyang You, Li Wenjie, et al. A Novel Composite Kernel Approach to Chinese Entity Relation Extraction[J]. Lecture Notes in Computer Science, 2009, 5459:236-247.

[8] Li Haiguang, Wu Xindong, Li Zhao, et al. A relation extraction method of Chinese named entities based on location and semantic features [J]. Applied Intelligence, 2013, 18(1): 1-15.

[9] Collins M, Duffy N. Covolution kernels for natural language[C]//Proceedings of the NIPS' 2001. Cambridge, MA 2001: 625-632.


[10] Zhang Ming, Zhang Jie, Su Jian, et al. A Composite Kernel to Extract Relations between Entities with both Flat and Structured Features[C]//Proceedings of the COLING-ACL'2006. Sydney, Australia, 2006: 825-832.

[11] Qian Longhua, Zhou Guodong, Zhu Qiaoming. Exploiting constituent dependencies for tree kernel-based semantic relation extraction[C]//Proceedings of the COLING'2008. Manchester, UK, 2008: 697-704.




郭剑毅(1964—),通信作者,硕士生导师,教授,主要研究领域为自然语言处理、信息抽取、机器学习等。

E-mail: gjade86@hotmail.com



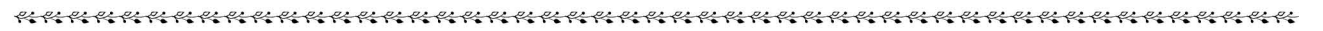
陈鹏(1987—),硕士研究生,主要研究领域为实体关系抽取。

E-mail: chen_peng0905@163.com



余正涛(1970—),教授,博士生导师,主要研究领域为自然语言处理、信息检索、机器翻译、机器学习等。

E-mail: ztyu@hotmail.com



(上接第 23 页)

[11] Witten I, Milne D. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links [C]//Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence. Chicago, USA, 2008: 25-30.


[12] Kim S, Toutanova K, Yu H. Multilingual named entity recognition using parallel data and metadata from Wikipedia[C]//Proceedings of ACL, Korea, 2012: 694-702.

[13] Han X, Zhao J. Structural semantic relatedness: a knowledge-based method to named entity disambiguation[C]//Proceedings of ACL, Sweden, 2010: 50-59.

[14] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]//Proceedings of WWW, Canada, 2007: 697-706.

[15] 叶正,林鸿飞,苏绥,等. 基于支持向量机的人物属性抽取[J]. 计算机研究与发展, 2007, 44: 271-275.

[16] 卢汉,曹存根,王石. 基于元性质的数量型属性值自动提取系统的实现[J]. 计算机研究与发展, 2010, 47(10): 1741-1748.




刘倩(1984—),博士,主要研究领域为自然语言处理、命名实体识别、网络文本挖掘、信息抽取。

E-mail: liuqian1104@126.com



刘冰洋(1987—),博士,主要研究领域为自然语言处理、命名实体识别、新词发现。

E-mail: liuctic@gmail.com



贺敏(1982—),博士,主要研究领域为自然语言处理、网络挖掘、信息安全。

E-mail: heminsmile@163.com