DOI:10.15986/j.1006-7930.2011.03.007

第43卷 第3期 西安建筑科技大学学报(自然科学版) Vol.43 No.3
2011年6月 J.Xi'an Univ. of Arch. & Tech.(Natural Science Edition) Jun. 2011

集成学习算法在实体关系抽取中的应用

董丽丽,高 山,张 翔

(1 西安建筑科技大学信息与控制工程学院,陕西 西安 710055;
2 西部建筑科技国家重点实验室(筹),陕西 西安 710055)

摘 要:针对基于特征向量的实体关系抽取方法中分类算法分类精度的不足,提出了基于集成学习算法的实体关系抽取方法.该方法将实体特征组合并转化为特征向量,使用集成学习中的ADABoost.MH算法来构造实体关系抽取的分类器,弱分类器采用决策树进行构造,通过提高分类效果好的分类器的权重和分类错误样本权重的方式来提高分类的精度,从而实现实体关系类别的识别.该方法在对《人民日报》语料库的测试中,得到了比较好的效果.

关键词:集成学习;实体关系抽取;特征向量;ADABoost.MH

中图分类号:TP391 文献标志码:A 文章编号:1006-7930(2011)03-0446-05

随着互联网的普及和发展,信息量正以指数规律飞速地增长,信息抽取[1](Information Extraction,IE)研究正是针对信息爆炸带来的问题而产生的.信息抽取目前的主要研究方向是命名实体识别、实体关系抽取和事件抽取.实体抽取确定了文章中的主要元素,而实体关系的抽取确定了实体之间的联系.实体关系抽取在数据结构化、信息检索和自动应答系统等领域有着重要的研究意义.

目前实体关系抽取的主要技术有基于知识库的方法和基于机器学习的方法.知识库的方法需要花费大量的时间和人力构建维护知识库,移植性差.为了弥补这一缺陷,泛化能力强的机器学习算法逐渐取代知识库方法成为目前实体关系抽取研究的主流.采用机器学习方法的实体关系抽取主要有基于特征向量的机器学习方法和基于核Kernel的机器学习方法,而基于核Kernel的机器学习方法不需要构造特征向量,但在计算关系之间距离的时候需要使用核函数,该方法的训练和预测速度太慢,不适合处理大量数据.因此目前多数对实体关系抽取的研究者采用基于特征向量的机器学习方法,该方法需要构造特征向量形式的训练数据,然后使用各种机器学习算法,如支持向量机(SVM)等作为学习机构造分类器,但一般分类器存在冗余特征以及过拟合问题,虽然SVM很好的解决了以上问题但SVM在大数据集上的训练熟练速度较慢,并且需要大量的存储资源和很高的计算能力成为了它的缺点[2].

本文针对上述问题,采用基于特征向量的机器学习方法,并选择集成学习方法中的ADABoost.MH算法作为分类方法,该方法克服了训练集的过拟合问题,而且它有很强的泛化能力,并且能够提高分类的正确率.论文的主要工作是首先找出一个句子中的全部实体对,然后根据实体和实体两边的上下文构造特征向量,最后通过分类器决定哪些是真正需要的实体关系.该方法在标注过的《人民日报》语料库上进行实验,得到了较好的实体关系抽取结果.

1 基于特征向量的实体关系抽取方法

基于特征向量的实体关系抽取主要思想是首先找到出现在文本中的所有可能存在关系的实体对,通过实体和实体的上下文将这些实体对构造为候选关系实例,并转化为机器可以识别的向量模型.然后通过机器学习算法对训练样本进行训练得到分类器,最后通过分类器对候选实例进行分类从而识别实体关系类别,并依据预定义的关系类别进行标注.

收稿日期:2010-10-20 修改稿日期:2011-04-12
基金项目:陕西省自然科学基金资助项目(2009JM8006);陕西省教育厅专项科研项目(2010JK620)
作者简介:董丽丽(1960-),女,福建福州人,教授,硕士生导师,主要研究领域为分布式系统与计算机网络应用、数据挖掘.

2　实体对的识别

实体关系抽取的首要任务是必须先识别实体, 这就需要对命名实体进行标注. 目前我们只考虑一个句子中的两个实体之间的关系, 而不考虑跨越句子的实体之间的关系, 因为在同一个句子中, 只有词语之间的距离在一定范围内的两个实体之间的关系才比较明确.

我们把出现在同一句子中, 并且两个实体的词语之间的距离在一定范围之内的两个实体称为一个实体对. 把出现在两个实体左右两边的 m 个词称为实体对的上下文. 例如: 某个已标注过的句子为 W_{Lm}, …, W_{L1}, E1, W_{mid1}, W_{mind2}, …, W_{midi}, E2, W_{R1}, …, W_{Rm}, 其中 $i \leqslant 2$ m, < E1, E2> 为一对实体对, 该实体对的上下文为:

$$W_{Lm}, \cdots, W_{L1}, W_{mid1}, W_{mid2}, \cdots, W_{midi}, W_{Rl}, \cdots, W_{Rm}$$

3　特征向量的构造

所谓特征向量是实例的一种数值化的表示方法, 通常使用一个词表作为特征向量的元素, 而向量中元素的值可以是 1 或 0(出现为 1, 否则为 0); 也可以是表示元素的近似程度值.

在特征向量构造方面, 本文采用了车万翔[3] 提出的特征向量构造方法. 该特征向量的构造如图 1 所示, 该向量的实验表明实体左右两边上下文窗口大小为 2 时, 分类算法效果最优.

E1. TYPE, E2. TYPE, E1. SUBTYPE, E2. SUBTYPE, ORDER,
W_{i-w}, W_{i-w-1}, …, W_{i-1}, W_{i+1}, …, W_{i+w-1}, W_{i+w}, t_{i-w}, t_{i-w-1}, …, t_{i-1}, t_{i+1}, …, t_{i+w-1}, t_{i+w},
W_{j-w}, W_{j-w-1}, …, W_{j-1}, W_{j+1}, …, W_{j+w-1}, W_{j+w}, t_{j-w}, t_{j-w-1}, …, t_{j-1}, t_{j+1}, …, t_{j+w-1}, t_{j+w}

图 1　特征向量

Fig. 1　Characteristic vector

其中, E.TYPE 为实体所属大类, E.SUBTYPE 为实体所属子类; Order 为实体对之间的位置关系, 分别标记为 0(表示 E1 出现在 E2 的左边), 1(表示 E1 出现在 E2 的右边). 而 i 和 j 分别为先后出现的两个实体的位置. W_k 和 t_k 分别为位置 k 处的汉语词和词性, 出现为 1, 不出现为 0[3] .

4　构建分类器

集成学习[5] (ensemble learning)的方法是近年来机器学习领域中一种流行的、用来提高学习精度的算法. 集成学习为解决同一个问题训练出多个分类器, 在对新的数据进行处理时, 将各个分类器的结论以某种方式进行组合. 这种方法能克服各个分类器对训练集的过拟合问题, 提高泛化能力, 从而更好的对新数据进行分类. 集成学习方法现在已应用到包括文本分类的多个领域, 并且被证明该方法优于某些单分类器方法[6] .

4.1　ADABoost.MH 算法

ADABoosT.MH 来源于 Boosting 家族的 ADABoost 算法, 是 Schapire 和 Singe 提出的一个专为解决文本分类问题的 BOOSTING 算法[6], 其主要思想是针对同一个训练集训练不同的弱分类器, 然后把这些弱分类器结果进行组合, 构成一个更强的强分类器. ADABoost.MH 算法是 ADABoost 算法的推广形式, 用来解决多类多标签问题, 并将多类转变为两类别分类问题, 利用 ADABoost 处理两类别分类问题的长处, 来提高分类的效果.

ADABoost 算法是将分类器定义为 $H(X) \rightarrow Y$, 自变量 X 所指的是特征空间, 而应变量 Y 则是{1, ……, k}代表 k 个类别. 而 ADABoost.MH 是把一个有 X 个实例, 并具有 k 种类别, 转化为有 $X \times Y$ 个实例, 而这 $X \times Y$ 个实例都只有两种类别, 这样就将样本集放大了 k 倍, 而将弱分类器定义为 h: $X \times Y \rightarrow R$ 形式, 并在 $X \times Y$ 上维持一个 $|X| \times |Y|$ 的权重分布, 分类器的输出值大于 0 表示支持 x_i 属于 y_i 类; 若输出值小于 0 表示支持 x_i 不属于 y_i 类.

4.2 算法描述

输入：训练集 $S=\{(x_1, Y_1), \cdots, (x_g, Y_g)\}$，其中 $x_n \in X$，$Y_j \subseteq Y=\{y_1, \cdots, y_m\}$，$(j=1, \cdots, g)$ 是类别集合，文本 x_j 属于它的每一个类别.

初始化 $$D_1(x_j, y_i)=\frac{1}{mg}\ 1\leqslant j\leqslant g,\ 1\leqslant i\leqslant m$$

训练过程 对 $t=1, \cdots, T$ 循环执行：

1）将分布 $D_t=(x_j, y_i)$ 输入弱假设学习器；

2）从弱学习器得到弱假设 $h_t: X\times Y\rightarrow R$ 其中 X 是所有可能的文本组成的集合；

3）选择参数 α_t；

4）修改错误实例的权值

$$D_{t+1}(x_j, y_i)=\frac{D_t(x_j, y_i)\exp(-\alpha_t Y_j[y_i]\, h_t(x_j, y_i))}{Z_t} \tag{1}$$

其中：$Z_t=\sum_{i=1}^{m}\sum_{j=1}^{g}D_t(x_j, y_i)\exp(-\alpha_t Y_j[y_i]\, h_t(x_j, y_i))$ 是一个归一化因子，保证 $\sum_{i=1}^{m}\sum_{j=1}^{g}D_{t+1}(x_j, y_i)=1$；若 y_i 在 Y_j 中出现，那么 $Y_j[y_i]=1$，否则 $Y_j[y_i]=-1$；$h_t(x_j, y_i)>0$，表示文本 x_j 属于类别 y_i；$h_t(x_j, y_i)<0$，则表示文本 x_j 不属于类别 y_i，而 $|h_t(x_j, y_i)|$ 表示该分类判定的可信度.

输出 最终假设：

$$h(x_j, y_i)=\sum_{t=1}^{T}\alpha_t h_t(x_j, y_i) \tag{2}$$

4.3 弱分类器的选择

ADABoost.MH 算法需要和一个比随机分类效果略好的弱分类器组成集成学习方法，以提高集成学习分类的效果，本文采用一阶决策树作为弱分类器，形式如式(3)所：

$$h_t(x_j, y_i)=\begin{cases} c_{0i} & \text{if} \quad w_{kj}=0 \\ c_{1i} & \text{if} \quad w_{kj}=1 \end{cases} \tag{3}$$

其中 w_{kj} 为第 k 个词在 j 类中是否出现，若出现则 $w_{kj}=1$，否则为 $w_{kj}=0$；c_{0i} 和 c_{1i} 是实数. 每次迭代中，出现的词以及 c_{0i} 和 c_{1i} 的值通常都是不同的.

$$c_{xi}=\ln\left(\frac{W_b^{xik}}{W_{-b}^{xik}}\right) \tag{4}$$

其中

$$W_b^{xik}=\sum_{j=1}^{g}D_t(x_j, y_i)\lambda(w_{kj}=x)\lambda(Y_j[y_i]=b). \tag{5}$$

式(5)中 $b\in\{1, -1\}$，$x\in\{0, 1\}$，$i\in\{1, \cdots, m\}$，为第 k 个出现的词，$\lambda(\pi)$ 表示特征函数(当 π 为真 $\lambda(\pi)=1$，否则 $\lambda(\pi)=0$).

5 性能评估指标及实验结果分析

实验数据使用的是《人民日报》标注语料，该语料是对《人民日报》1998年上半年的纯文本语料进行了词语的切分和词语类型标注制作而成的.

5.1 性能评估指标

实验对以下两种方法做了比较，方法1是采用SVM 分类的基于特征向量的方法，方法2是本文介绍的方法，使用了集成学习算法. 在实验中，我们选取ACE 会议定义的两大类及4个子类作为预定义关系类别，这些类别包括Org-AFF 及其Employment 和Sports-Affiliation 子类；PART-WHOLE 及其Artifact 和Geographical 子类. 实验系统从标注过的语料文本中自动生成出5512个关系实例，随机的选择其中的1/4作为测试数据，其余的3/4作为训练数据.

实验的评测标准采用 F 值对最终系统的性能进行评价，其定义如式(6)：

$$F=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}} \tag{6}$$

其中准确率(Precision)、召回率(Recall)的定义如式(7)和式(8)为:

$$\text{Precision}=\frac{\text{某类被正确分类的实例个数}}{\text{分类器预测的某类实例总数}} \tag{7}$$

$$\text{Recall}=\frac{\text{某类被正确分类的实例个数}}{\text{测试数据中某类实例总数}} \tag{8}$$

5.2　实验结果分析

实验结果见表 1 和表 2,表 1 列举了 4 个子类关系抽取的结果;表 2 列举了两种分类方法关于两大类关系抽取的结果.

表 1　两大类关系中 4 种子类关系抽取结果

Tab. 1　The result of four sub-category in two kinds of relationship

Type	Subtype	P/%	R/%	F/%
ORGAFF	Employment	82.21	79.69	80.93
	Sport Affiliation	80.72	77.97	79.32
PART	Artifact	78.14	76.11	77.11
WHOLE	Geographical	81.27	78.69	79.96

表 2　SVM 和 ADABoost.MH 方法对比

Tab. 2　Comparison of SVM and ADABoost.MH methods

Approaches	P/%	R/%	F/%
SVM	79.11	76.89	77.98
ADABoost.MH	80.59	78.12	79.33

从结果可以看出基于集成学习算法的实体关系抽取是可行的,但系统抽取结果中仍存在一些问题,一部分错误是由于词性标注模块和实体识别模块未能正确标注引起的,一部分错误由于分类器本身引起的错误,例如"张三 2004 年毕业于北京大学英语系后留校任教"中包含两个关系,一个是"张三"毕业于"北京大学",一个是"张三"在"北京大学"工作,目前只能抽取第 1 个实体关系,而隐含的在"北京大学"任教的关系没有正确抽取.

6　结　语

本文介绍了基于集成学习算法的实体关系抽取方法,通过构造特征向量,使用 ADABoost.MH 算法和决策树结合的集成学习算法,通过对 ACE 定义的两大类中 4 个子类的实体关系抽取,实验证明,集成学习算法提高了实体关系抽取的精度.本研究的进一步工作是的特征向量表示方法进行改进从而提高实体关系抽取准确率.

参考文献　References

[1] 程显毅, 朱　倩, 王　进. 中文信息抽取原理及应用[M]. 北京. 科学出版社, 2010: 70.
CHENG Xian-yi, ZHU Qian, WANG Jin. Chinese information extraction principle and application[M]. Beijing,. Science Publishing Company. February, 2010: 70.

[2] 苏金树, 张博峰, 徐昕. 基于机器学习的文本分类技术研究进展[J]. Journal of Software, 2006, 17(9): 1848-1859.
SU Jin-Shu, ZHANG Bo-Feng, XU Xin. Advances in Machine Learning Based Text Categorization[J]. Journal of Software. 2006, 17(9): 1848-1859.

[3] 车万翔, 刘　挺, 李　生. 实体关系自动抽取[J]. 中文信息学报, 2004, 19(2): 2.
CHE Wan-xiang, LIU Ting, LI Sheng. Automatic Entity Relation Extraction[J]. Journal of Chinese information processing, 2004, 19(2): 2.

[4] ACE. 2007. The nist ace evaluation website. [OL]. [2010/8/27]. http://www.nist.gov/speech/tests/ace/ace07/.

[5] LIANG YingYi. Integrated learning review[OL]. http://soft.cs.tsinghua.edu.cn/~keltin/docs/ensemble.pdf

[6] SCHAPIRE R, SINGER Y. BoosTexter: a boosting based system for text categorization[J]. Machine Learning, 2000, 39(203): 135-168.

[7] ZHOU GuoDong, SU Jian, ZHANG Jie, et al. Exploring Various Knowledge in Relation Extraction[J]. Association for Computational Linguistics, 2005: 427-434.

[8] 姚谦峰, 侯莉娜, 黄　炜. 给予遗传算法的多层密肋壁板结构优化设计方法研究[J]. 西安建筑科技大学学报: 自然科学版, 2009, 41(4): 445-460.
YAO Qian-feng, HOU Li-na, HUANG Wei. Optimization design method of multi-storied multi-ribbed slab structure based on GA[J]. J. Xi'an Univ. of Arch. & Tech.: Natural Science Edition, 2009, 41(4): 455-460.

Application of the research on extraction of entity relationship based on integrated learning algorithm

DONG Li-li, GAO Shan, ZHANG Xiang

(1 School of information and control engineering, Xi'an Univ. of Arch. & Tech., Xi'an 710055, China;
2 State Key Laboratory of Architecture Science and Technology in West China(XAUAT), Xi'an 710055, China)

Abstract: To overcome the classification accuracy defects of traditional classification algorithm, a method of integrated learning is brought forward. The method which combined entity characteristics and translated entity characteristics into feature vector introduced an integrated learning algorithm. ADABoost. MH algorithm is used to divide weak classifier. By improving the weight of good classifier and wrong results to increase classification accuracy realized the recognized classes of entity. The method proved to be effective in test of the corpus of the people's Daily.

Key words: *integrated learning*; *extraction of entity relationship*; *feature vector*; *adaboost. mh*

Biography: DONG Li-li, Professor, Xi'an 710055, P. R. China, Tel: 0086-13572269987, E-mail: donglilixjd@163. com

(上接第421页)

[8] 孙 剑. 中国建筑业供求关系监测模型研究[J]. 建筑经济, 2005(7): 33-37.
SUN Jian. Monitoring model to study the relation between supply and demand of China construction[J]. Construction economy, 2005(7): 33-37.

[9] 孙延芳. 我国建筑业周期波动的测度[J]. 西安建筑科技大学学报: 自然科学版, 2010, 42(4): 596-598.
SUN Yan-fang. Measuring the cyclic fluctuations in China s construction industry [J]. J. Xi'an University of Architecture & Technology: Natural Science Edition, 2010, 42(4): 596-598.

Monitoring model on the supply & demand balance of housing provident fund market

SONG Jin-zhao, LIU Xiao-jun, DONG Hong-liang

(School of Management, Xi'an Univ. of Arch. & Tech., Xi'an 710055, China)

Abstract: The relationship between supply and demand affects the development of the housing provident fund market deeply. By comprehensive analysis, this paper finds some key factors of the supply and demand of the housing provident fund market, and sets up a model to monitor their change according to the intensity of supply and demand. Then, by an empirical analysis, this paper monitors the balance status of Shanghai housing provident fund market. As its result stays in line with the actual situation, the model has a higher reliability.

Key words: *the housing provident fund*; *effective balance of supply and demand*; *supply index*; *demand index*; *balance index*

Biography: SONG Jin-zhao, lecturer, Candidate for Ph. D., Xi'an 710055, P. R. China, Tel: 0086-29-13571851316, E-mail: songjinzhao78@163. com