基于三支决策的两阶段实体关系抽取研究

朱艳辉[1,2],李 飞[1,2],胡骏飞[1,2],钱继胜[3],王天吉[1,2]
ZHU Yanhui[1,2], LI Fei[1,2], HU Junfei[1,2], QIAN Jisheng[3], WANG Tianji[1,2]

1.湖南工业大学 计算机学院,湖南 株洲 412008
2.湖南工业大学 湖南省智能信息感知及处理技术重点实验室,湖南 株洲 412008
3.中国人民银行 铜陵市中心支行,安徽 铜陵 244000
1.School of Computer Science, Hunan University of Technology, Zhuzhou, Hunan 412008, China
2.Hunan Key Laboratory of Intelligent Information Perception and Processing Technology, Hunan University of Technology, Zhuzhou, Hunan 412008, China
3.The People's Bank of China Tongling Central Sub-branch, Tongling, Anhui 244000, China

ZHU Yanhui, LI Fei, HU Junfei, et al. Research on two-stage entity relation extraction based on three-way decisions. Computer Engineering and Applications, 2018, 54(9):145-150.

Abstract: As one of the important research topics in information extraction, entity relationship extraction is of great significance to the construction of knowledge graph data layer. This paper proposes a two-stage classification technique based on three-way decisions to extract the entity relationship. Firstly, the SVM three-decisions classifier is constructed to implement the first phase entity relation extraction. The softmax multi-class function is used as a probability function of three-way decisions, Then, the *K*NN classifier is used to classify the three-way decisions middle domain sample into two-stage classification. According to the corpus of ACE2005 as the experimental data, the results of the three-way decisions two-stage classification are compared with the traditional SVM method. The experimental results show that the two-stage entity relation extraction method based on three-way decisions has achieved good classification effect.

Key words: entity relation extraction; three-way decisions; Support Vector Machine(SVM); *K*-Nearest Neighbor(KNN); softmax function

摘 要: 实体关系抽取作为信息抽取研究的重要研究课题之一,对知识图谱数据层的构建有着重要的意义。提出一种基于三支决策的两阶段分类技术实现实体关系抽取,首先构建SVM三支决策分类器实现第一阶段实体关系抽取,采用softmax多分类函数作为三支决策概率函数,然后采用*K*NN分类器对三支决策分类后的中间域样本进行二阶段分类。以ACE2005的语料作为实验数据,将三支决策两阶段分类结果与传统SVM方法分类结果进行比较,实验结果表明,基于三支决策的两阶段实体关系抽取方法取得了很好的分类效果。

关键词: 实体关系抽取;三支决策;支持向量机(SVM);*K*最近邻(*K*NN);softmax函数
文献标志码: A **中图分类号:** TP391 doi:10.3778/j.issn.1002-8331.1710-0153

1 引言

随着计算机的普及和知识工程的蓬勃发展,信息量正以指数级的规模爆炸式增长。人们迫切地需要一些自动化的工具帮助人们在海量数据源中快速检索到需要的知识。信息抽取(Information Extraction)研究以此为背景应运而生。其主要目的是将互联网中海量的非结构化数据转化为结构化或半结构化信息形成"知识",供用户查询以及进一步分析挖掘。信息抽取在信息检

基金项目: 国家自然科学基金(No.61402165);模式识别国家重点实验室开放课题(No.201700009);湖南省教育厅重点项目(No.15A049);湖南工业大学重点项目(No.17ZBLWT001KT006);湖南省研究生创新基金(No.CX2017B688)。

作者简介: 朱艳辉(1968—),女,教授,计算机学会高级会员,研究领域为文本处理和知识工程;李飞(1992—),男,硕士研究生,研究领域为自然语言处理,E-mail:flytoskye@163.com;胡骏飞(1990—),男,硕士研究生, 研究领域为智能信息处理;钱继胜(1982—),男,工程师,研究领域为金融信息化和知识工程;王天吉(1985—),男,硕士研究生,研究领域为文本处理。

收稿日期: 2017-10-18 **修回日期:** 2017-12-01 **文章编号:** 1002-8331(2018)09-0145-06

索、知识表示、篇章理解、智能问答等领域具有重要的应用价值。信息抽取从文本中抽取出特定的事实称之为“实体”,然而大多数应用中不仅需要“实体”,还要确定这些实体的关系,称其为实体关系抽取。美国国家标准技术研究院(NIST)组织了自动内容抽取(Automatic Content Extraction,ACE),其评测任务之一就是实体关系识别。实体关系抽取通过识别命名实体,进而抽象出实体间关系类型,如NIST定义了制造使用关系(ART)、组织机构从属关系(ORG-AFF)、局部整体关系(PART-WHOLE)等关系类型。因此可将关系抽取问题转化为多分类问题。首先识别出句子中所有的实体对,然后使用分类器决定实体关系类型属于预定义的哪一类。

许多学者采用SVM分类器进行实体关系抽取,车万翔[1]等人使用SVM构造不同窗口大小的特征向量,在ACE2004语料上取得了较好的分类效果。刘绍毓[2]等对SVM模糊边界样本进行双投票,对模糊样本采用*K*NN分类器进行二次分类,大大提高了实体关系抽取的准确率。但是,虽然SVM具有较强的抗噪声能力和较高的分类准确率等优点,但该分类器对于分布在超平面附近区域的样本分类效果不理想。当处理多分类任务时,样本在超平面附近的类交叠区域的分类效果更差。故随着分类类别数的增多,由于各个类别样本交叠愈加严重,从而影响分类准确率。

三支决策理论[3-6]是传统二支决策理论的拓展,二支决策只考虑接受或者拒绝(或者是或否)两种选择。但是实际应用中,由于信息的不确定性和不全面性,无法明确对一个事物明确的判断接受或是拒绝。因此,Yao(姚一豫)[7-8]提出了三支决策理论,当判决信息不足以判断接受或者拒绝时,采用不承诺选择,然后再加入细粒度信息进行下一步判断[9]。李金海[10]论述了三支决策与概念格相结合的研究进展,针对两个结合点:三支概念分析和三支概念学习进行对比分析,提出了两种思维的互补性。并且提出一种建立不完整的上下文近似概念格的新方法[11],通过从不完全决策环境中提取非冗余近似决策规则,进一步提高了三支决策在信息不完备情况下的决策效率。二支决策和三支决策就应用场景而言各有优劣,在信息充足、消息准确时,采用二支决策,可使得决策迅速简洁。在信息不足或者获取信息代价过大时,适合使用三支决策,可以权衡利弊,等待细粒度的信息,再做出进一步判断。三支决策策略提供了一个很好的权衡资源和效益的决策框架。

本文将三支决策应用到实体关系抽取领域,对信息不足以判断实体关系属于哪一类型的样本,引入中间类别(中间域)。针对SVM分类器交叠区域样本难以界定的问题,提出一种基于三支决策的两阶段实体关系抽取方法。首先构建SVM三支决策分类器实现第一阶段实体关系抽取,采用softmax函数作为三支决策概率函数,然后采用*K*NN分类器对三支决策分类后的中间域样本进行二阶段分类。并将结果与SVM分类方法和一阶段SVM三支决策分类方法进行比较实验,实验结果表明,基于三支决策两阶段分类实体关系抽取方法取得了很好的抽取效果。

2 三支决策理论

三支决策理论是在粗糙集和决策粗糙集理论之上提出的,Yao通过对粗糙集理论中的正、负、边界区域语义方面研究,提出了从三支决策角度解释粗糙集中规则提取问题。其规则分别对应对象所属的正、负、边界三个区域,根据对象所属区域不同,分别判决该对象属于目标类、不属于目标类、不承诺是否属于目标类的三支决策策略,对于决策粗糙集模型所需的阈值参数可由决策损失函数决定。

2.1 决策粗糙集理论

定义一个四元组 $W=(U,At=B\bigcup C,\{V_a|a\in At\},\{I_a|a\in At\})$,其中 U 是一个有限且非空的数据对象集合[12], At 是一个非空且有限属性集合, B 是条件属性, C 是决策属性, $B\bigcap C=\varnothing$, V_a 为属性值的集合, I_a 是对象 U 到 V_a 的一个映射,称为信息函数,即将集合 U 映射到属性值域 V_a 上。(U,E_A) 是属性集合 A 上的近似集合, U/E_A 是基于关系集合 E_A 对对象集合的划分, E_A 定义如下:

$$E_A=\{(x,y)\in U\times U|\forall a\in A\subseteq C\subset At,I(a)=I_a(y)\} \tag{1}$$

则包含对象 x 的等价类可表示为:

$$[x]_A=[x]=\{y\in U|(x,y\in E_A)\} \tag{2}$$

判断一个对象是否属于决策类可用状态集合 $\Omega=\{X,\neg X\}$ 表示,则等价类 $[x]$ 属于决策类 X 的概率函数为:

$$p(X|[x])=\frac{|X\bigcap[x]|}{|[x]|} \tag{3}$$

不属于决策类 X 的概率函数为:

$$p(\neg X|[x])=\frac{|\neg X\bigcap[x]|}{|[x]|}=1-p(X|[x]) \tag{4}$$

2.2 三支决策阈值

Yao等人提出了决策粗糙集模型,并定义了如下三个域(设阈值 $0\leqslant\beta<\alpha\leqslant1$):

$$\begin{cases}POS_{(\alpha,\beta)}(X)=\{x\in U|P(X|[x])\geqslant\alpha\}\\BND_{(\alpha,\beta)}(X)=\{x\in U|\alpha<P(X|[x])<\beta\}\\NEG_{(\alpha,\beta)}(X)=\{x\in U|P(X|[x])\leqslant\beta\}\end{cases} \tag{5}$$

其中 $POS_{(\alpha,\beta)}(X)$、$BND_{(\alpha,\beta)}(X)$、$NEG_{(\alpha,\beta)}(X)$ 分别称为 X 的正域、边界域、负域。

当对象 x 属于决策类 X 时,令 λ_{pp}、λ_{np}、λ_{bp} 为分别划分到 $POS_{(\alpha,\beta)}(X)$、$BND_{(\alpha,\beta)}(X)$、$NEG_{(\alpha,\beta)}(X)$ 的损

失函数。当对象 x 不属于决策类时，则令 λ_{pn}、λ_{bn}、λ_{nn} 为划分到相同三个域的损失函数。则损失函数表如表1所示。

表1　损失函数表

类型	$POS_{(\alpha,\beta)}(X)$	$BND_{(\alpha,\beta)}(X)$	$NEG_{(\alpha,\beta)}(X)$
属于决策类	λ_{pp}	λ_{np}	λ_{bp}
不属于决策类	λ_{pn}	λ_{bn}	λ_{nn}

对于三个域的风险决策，结合贝叶斯决策理论给出的最小风险决策规则。可知：

$$\begin{cases}\alpha=\dfrac{(\lambda_{pn}-\lambda_{bn})}{(\lambda_{pn}-\lambda_{bn})+(\lambda_{bp}-\lambda_{pp})}\\ \gamma=\dfrac{(\lambda_{pn}-\lambda_{nn})}{(\lambda_{np}-\lambda_{pp})+(\lambda_{pn}-\lambda_{nn})}\\ \beta=\dfrac{(\lambda_{bn}-\lambda_{nn})}{(\lambda_{bn}-\lambda_{nn})+(\lambda_{np}-\lambda_{bp})}\end{cases}\tag{6}$$

则以上决策规则简化如下：

$$\begin{cases}\text{正规则：}P(X[\![x]\!]_A)\geqslant\alpha,[x]_A\subseteq POS_{(\alpha,\beta)}(X)\\ \text{负规则：}P(X[\![x]\!]_A)\leqslant\beta,[x]_A\subseteq NEG_{(\alpha,\beta)}(X)\\ \text{边界规则：}\beta<P(X[\![x]\!]_A)<\alpha,[x]_A\subseteq BND_{(\alpha,\beta)}(X)\end{cases}\tag{7}$$

在 $[x]_A$ 的情况下，如果 X 发生的概率大于等于 α，则将 $[x]_A$ 划分为 X 的正域，如果 X 发生概率大于 β 小于 α，则将 $[x]_A$ 划分为 X 的边界域，如果 X 发生的概率小于等于 β，则将 $[x]_A$ 划分为 X 的负域[13]。

3　特征抽取

本文采用词汇、实体类型、位置等作为文本特征。

(1)词汇

实体本身所包含的所有词汇，以及实体左右的词汇对确定实体之间的关系有很好的作用。例如，“微软公司创始人比尔盖茨从哈佛大学退学后创办微软公司”。实体“微软”和实体“比尔盖茨”属于雇佣关系，其中在实体“微软”附近的词(公司、创办)对实体“比尔盖茨”很有指示作用。所以实体窗口词对于分类也十分关键，但是窗口太大，会引入太多无关信息。窗口太小，又会导致重要信息的遗漏。车万翔等人经过重复实验验证了在窗口取2时，分类能取得最好的效果，故本文取实体上下文窗口为2的词，如表2所示。E1、E1pos表示实体1词汇及词性，E2、E2pos表示实体2词汇及词性。E1L1、E1L1pos表示实体左侧第一个词及其词性，E1L2、E1L2pos表示实体左侧第二个词及其词性，E1R1、E1R1pos、E1R2、E1R2pos表示实体右侧第一、二个词及其词性。E2同理。

(2)位置特征

实体的位置特征以及实体的先后顺序对于关系类型有很大影响。董静[14]等人对ACE语料样本特征进行分析，提取实体包含和非包含关系特征对实体关系抽取，证明了实体包含和非包含特征对实体关系抽取有一定影响。本文采取的实体位置特征如表3所示。

表2　实体词和上下文特征

词	词性	实体特征
微软	NI	E1
公司	N	E1R1(E2L2)
创始人	N	E1R2(E2L1)
比尔盖茨	NH	E2
从	P	E2R1
哈佛	NS	E2R2
大学	N	
退学	NV	
后	ND	E1L2
创办	V	E1L1
微软	NI	E1
公司	N	E1R1

表3　实体位置特征

特征	描述
0	实体E1位于E2左侧
1	实体E1位于E2右侧
2	实体E1包含实体E2

(3)实体类型

实体关系分类中实体类型及其组合特征[15]是一个非常重要的特征，对分类准确与否至关重要，实体类型特征标记如表4所示。

表4　实体类型特征

特征	描述
E1TYPE	实体E1的类型
E1SUBTYPE	实体E1的子类型
E2TYPE	实体E2的类型
E2SUBTYPE	实体E2的子类型
ETYPEC1	实体E1类型和子类型的组合特征
ETYPEC2	实体E2类型和子类型的组合特征

4　基于三支决策的两阶段实体关系抽取

本文通过构造SVM三支决策分类器，进行一阶段实体关系抽取，然后采用*K*NN分类器对三支决策中间域样本进行二阶段实体关系抽取，从而实现基于三支决策的两阶段实体关系抽取。实体关系抽取流程图如图1所示。

4.1　SVM三支决策分类器构建

鉴于实体关系抽取是一个多分类问题，SVM提供了多分类方法：一种是one-against-rest方法，基本思想是对于 $M(M\geqslant3)$ 类样本，将其中一类和其余类分别作为正、负例来训练分类器，M 个类别需构建 M 个分类器。另一种是one-against-one方法，基本思想是对于 $M(M\geqslant3)$ 类样本，每两类训练一个分类器，M 个类别

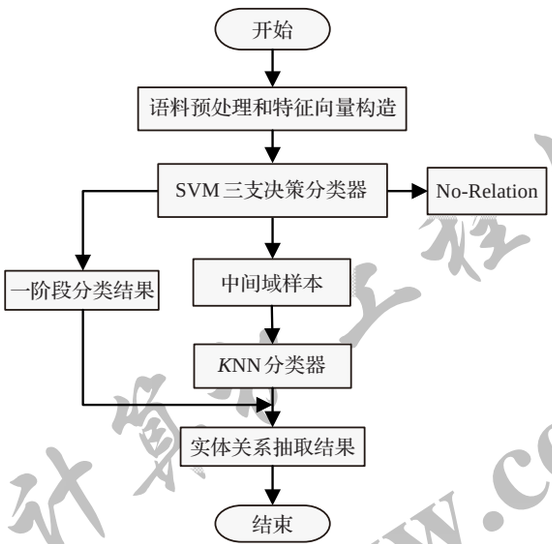

图1 实体关系抽取流程图

需构建 $M(M-1)/2$ 个分类器。鉴于one-against-rest方法分类速度较快,训练分类器数目较少,本文采用one-against-rest方法。在三支决策分类器的构建中,针对多分类问题,采用softmax函数作为概率函数,计算每个样本属于某类的概率值,计算公式如式(8)所示:

$$\sigma(z(x_i))=\frac{e^{z(x_i)}}{\sum_{k=1}^{k}e^{z(x_k)}},i=1,2,\cdots,k \tag{8}$$

其中 $z(x_i)=\sum_{i\in SV}\alpha_i y_i k(x,x_i)+b$, x_i 和 $\alpha_i y_i$ 是由SVM所确定的支持向量及其参数, b 是SVM分类器定义的超平面参数[16]。其中SVM核函数采用径向基核函数(RBF-Kernel):

$$k(x,x_i)=e^{-r<x_i-x,x_i-x>2} \tag{9}$$

SVM三支决策分类器构造算法如下:

输入:训练集 U ,测试集 C ,类别集合 k 。

输出:实体类别集 $Set(T)$,边界域(中间域)样本集 $Set(MID)$,No-Relation样本集 $Set(F)$ 。

训练阶段:

步骤1 输入训练集样本集合 U 。

步骤2 使用SVM分类器对训练集 U 进行训练,得到SVM分类模型。

测试阶段:

步骤1 输入测试集样本集合 C 。

步骤2 for $c_i\in C$,使用训练好的SVM分类器进行分类。

步骤3 由式(8)计算 C 中所有样本对象分别属于类别集合 k 中某类的概率,并构成概率矩阵集合 P 。

步骤4 if $\sigma(z(c_i))\geqslant\alpha$,样本 $c_i\rightarrow POS_{(\alpha,\beta)}(X)$,将 c_i 加入 $Set(T)$ 。

步骤5 else if $\beta<\sigma(z(c_i))<\alpha$,样本 $c_i\rightarrow BND_{(\alpha,\beta)}(X)$,将 c_i 加入 $Set(MID)$ 。

步骤6 else if $\sigma(z(c_i))\leqslant\beta$,样本 $c_i\rightarrow NEG_{(\alpha,\beta)}(X)$,将 c_i 加入 $Set(F)$ 。

步骤7 end。

由算法可以看出,首先对 n 个样本进行分类,并且要计算 n 个样本分别属于类别集合 k 中某类的概率,故算法需执行 $n\times k$ 次,由于 k 为常数,所以时间复杂度与 n 成线性关系, $T(n)=O(n)$,算法从时间复杂度的角度分析是有效的。

4.2 SVM三支决策分类器阈值计算

对于阈值 α 与 β ,作如下假设:

$$\begin{cases}\lambda_{pp}=\lambda_{nn}=0;\lambda_{np}=\eta\lambda_{bp}\\ \lambda_{pn}=\eta\lambda_{bn};\lambda_{bn}=\lambda_{bp}\end{cases} \tag{10}$$

则由式(6)和(10)可得:

$$\begin{cases}\alpha=\dfrac{(\lambda_{pn}-\lambda_{bn})}{(\lambda_{pn}-\lambda_{bn})+(\lambda_{bp}-\lambda_{pp})}=\dfrac{\eta-1}{\eta}\\ \beta=\dfrac{(\lambda_{bn}-\lambda_{nn})}{(\lambda_{bn}-\lambda_{nn})+(\lambda_{np}-\lambda_{bp})}=\dfrac{1}{\eta}\end{cases} \tag{11}$$

由于 $\alpha>\gamma>\beta$,所以 $\eta>2$, η 的最后取值由实验结果确定[17]。

4.3 基于KNN的三支决策中间域样本二阶段分类

KNN算法是一种简单易行的无参数分类方法,该算法对非正态分布的数据具有较高的分类准确率,具有鲁棒性强、易于实现等优点,在人工智能领域、模式识别等领域已经取得广泛的应用[18]。但该算法时间及空间复杂度随着样本集合增大而增高,由于中间域样本数较少,故本文选用KNN算法在第二阶段对中间域样本进行二次分类。该算法基本思路是:如果某样本在特征空间的K个最相似的样本中的大多数属于某类别,则该样本也属于该类别。本文采用KNN分类器作为二阶段分类器,对三支决策中间域样本集 $Set(MID)$ 进行二次分类,使用欧式距离计算样本间距离:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{12}$$

其中 X 与 Y 分别表示样本集中某两样本构成的特征向量 $X=(x_1,x_2,\cdots,x_n)$ 和 $Y=(y_1,y_2,\cdots,y_n)$, $d(X,Y)$ 表示两样本之间距离。

5 实验设计与结果分析

5.1 实验数据选取及预处理

本文实验语料采用ACE2005中文评测语料,数据来源为广播新闻(Broadcast News),新华社新闻(XinHua News)[19]。并选取前8 000篇作为训练语料,后1 317篇作为测试语料。ACE的训练数据,不仅标注实体以及实

体的属性，还详细标注了实体关系以及关系的属性，数据以及标注结果以XML格式存储，句子中任意两个实体之间即形成一个实例，表5列出了本文所选取语料所有实例的统计情况。

表5　实例统计信息

类别	个数	类别	个数
ART	520	PART-WHOLE	1 938
GEN-AFF	1 694	PER-SOC	498
METONYMY	39	PHYS	1 408
ORG-AFF	1 927	TOTLE	8 024

由表5可知，转喻关系(METONYMY)类型仅占39个，且转喻关系类型不包含任何子类型，故本实验剔除转喻关系(METONYMY)类型，只考虑除METONYMY(转喻关系)外的6类关系类型。

语料预处理包括分词、词性标注等。分词采用Python自带的jieba分词，抽取样本集中所有实体词汇，构成实体词典，作为jieba分词的自定义词典，大大避免了实体词汇被错分的情况。词性标注采用jieba自带的词性标注工具[15]。

5.2　评价标准

本文采用信息检索的通用评价方法，准确率(P)、召回率(R)和F值定义如下：

$$\begin{cases} P=\dfrac{\text{某类被正确分类的个数}}{\text{分类器预测的某类总数}} \\ R=\dfrac{\text{某类被正确分类的个数}}{\text{测试数据中某类总数}} \\ F=\dfrac{2\times P\times R}{P+R} \end{cases} \tag{13}$$

对两阶段分类结果进行加权处理作为最终分类结果。公式如下：

$$\begin{cases} P_{\text{Final}}=\dfrac{A1}{ALL}P_1+\dfrac{M1}{ALL}P_2 \\ R_{\text{Final}}=\dfrac{A1}{ALL}R_1+\dfrac{M1}{ALL}R_2 \\ F_{\text{Final}}=\dfrac{A1}{ALL}F_1+\dfrac{M1}{ALL}F_2 \end{cases} \tag{14}$$

其中，ALL为所有样本数，$A1$为一阶段中分到各实体类别的的样本总数，$M1$为一阶段中间域样本数。P_1、R_1、F_1分别为一阶段准确率、召回率和F值。P_2、R_2、F_2分别为二阶段准确率、召回率和F值。

5.3　实验结果分析

5.3.1　参数η取值实验

对参数η进行取值实验，实验区间为[2.0，4.0]，实验结果如图2所示。

从图2~7可得出如下结论：随着η值的增大，准确率逐渐上升，F值在[1.7，2.9]之间逐渐增大，而召回率在[1.7，2.9]区间缓慢下降，在2.9以后呈直线下降，η取均值2.85时准确率、召回率、F值达到最高。取η=2.85时，实验结果如表6所示。

表6　一阶段SVM三支决策分类结果(η=2.85)

关系类型	P	R	F
ART	1.00	0.38	0.55
GEN-AFF	0.77	0.75	0.76
ORG-AFF	0.83	0.93	0.88
PART-WHOLE	0.81	0.85	0.83
PER-SOC	0.72	0.98	0.83
PHYS	0.85	0.78	0.81
Average	0.82	0.80	0.81

5.3.2　二阶段中间域样本*KNN*实验

由上节可知，η取2.85时效果最好，故下面实验取η=2.85，将其代入式(11)，可得：

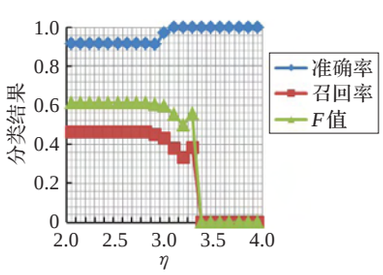

图2　类别Ⅰ参数η取值实验

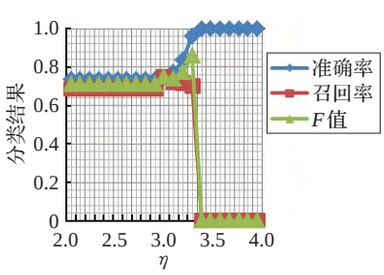

图3　类别Ⅱ参数η取值实验

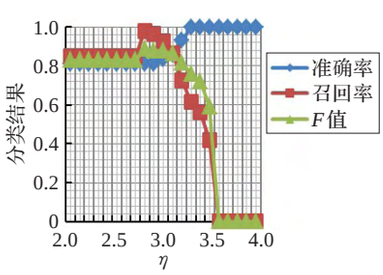

图4　类别Ⅲ参数η取值实验

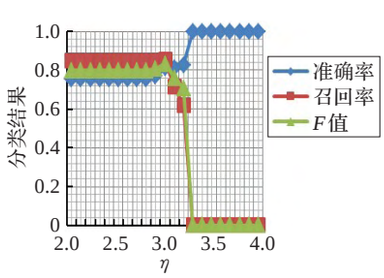

图5　类别Ⅳ参数η取值实验

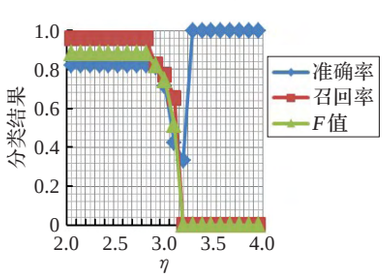

图6　类别Ⅴ参数η取值实验

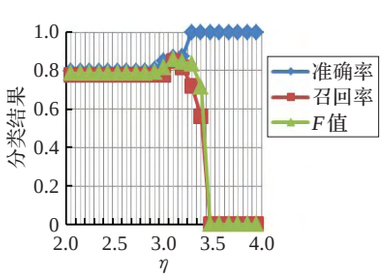

图7　类别Ⅵ参数η取值实验

$$\begin{cases}\alpha=\dfrac{(\lambda_{pn}-\lambda_{bn})}{(\lambda_{pn}-\lambda_{bn})+(\lambda_{bp}-\lambda_{pp})}=\dfrac{\eta-1}{\eta}=\dfrac{2.85-1}{2.85}=0.65\\ \beta=\dfrac{(\lambda_{bn}-\lambda_{nn})}{(\lambda_{bn}-\lambda_{nn})+(\lambda_{np}-\lambda_{bp})}=\dfrac{1}{\eta}=\dfrac{1}{2.85}=0.35\end{cases}\tag{15}$$

将中间域样本 *Set*(*MID*) 输入训练好的 *K*NN 分类器中进行实体关系抽取。实验结果如表7所示。

表7 二阶段 *K*NN 实体关系抽取实验结果

关系类型	*P*	*R*	*F*
ART	1.00	0.75	0.86
GEN-AFF	0.79	0.74	0.76
ORG-AFF	0.89	0.90	0.89
PART-WHOLE	0.81	0.95	0.87
PER-SOC	0.89	1.00	0.94
PHYS	0.94	0.84	0.89
Average	0.87	0.86	0.86

5.3.3 实验结果对比

选择效果最好的 η =2.85的两阶段分类加权平均实验结果与一阶段SVM三支决策分类结果、文献[1]中结果进行比较。结果如表8所示。

表8 本文方法与各方法结果比较

分类结果	*P*	*R*	*F*
文献[1]结果(传统SVM结果)	0.76	0.70	0.73
一阶段SVM三支决策分类结果	0.82	0.80	0.81
三支决策两阶段加权分类结果	0.85	0.81	0.82

由表8可知,一阶段SVM三支决策分类结果较传统SVM分类结果提升效果较为显著,这表明三支决策方法在实体关系抽取领域的应用是有效的。基于三支决策两阶段分类(本文方法)结果相较于传统SVM分类结果在准确率、召回率、*F* 值上分别提高了9%、11%、9%,表明本文方法大大提高了实体关系抽取的效果,而三支决策两阶段分类结果相较于一阶段SVM三支决策分类结果也有一定的提升,证明了使用*K*NN分类器对中间域样本的处理对提高实体关系抽取效果也是有效的。

6 总结与展望

本文以ACE2005中文评测语料进行研究,提出了一种基于三支决策的SVM-*K*NN两阶段实体关系抽取方法。实验结果表明,该方法有效提高了实体关系抽取的分类效果。本文研究还存在一些不足之处:(1)三支决策的损失函数、阈值仅根据专家经验进行了简单预设;(2)文本特征选择还偏于简单,应研究更细粒度的特征如语义特征、句法路径特征、包含非包含特征等。接下来的工作,将对上述不足之处进行进一步探讨,以进一步提高实体关系的抽取效果。

致谢 本文研究内容得益于作者朱艳辉在加拿大Regina大学访学期间来自于姚一豫教授的悉心指导,在此对姚一豫教授表示深深的感谢。

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] 刘绍毓,周杰,李弼程,等.基于多分类SVM-KNN的实体关系抽取方法[J].数据采集与处理,2015,30(1):202-210.

[3] Pawlak Z.Rough sets[J].International Journal of Computer and Information Sciences,1982,11(5):341-356.

[4] Pawlak Z.Roughset:Theoretical aspects of reasonsing about data[M].Dordrecht:Kluwer Academic Publishers,1991.

[5] Yao Y Y,Wong S K M,Lingras P.A decision-theoretic rough set model[C]//The 5th International Symposium on Methodologies for Intelligent Systems,1990.

[6] Yao Y Y,Wong S K M.A decision theoretic framework for approximating concepts[J].International Journal of Man-Machine Studies,1992,37:793-809.

[7] Yao Y Y.An outline of a theory of three-way decisions[C]//Proceedings of the 8th International RSCTC Conference,2012.

[8] Yao Y Y.The superiority of three-way decisions in probabilistic rough set models[J].Information Sciences,2011,181:1080-1096.

[9] 张燕平,邹慧锦,邢航,等.CCA三支决策模型的边界域样本处理[J],计算机科学与探索,2014,8(5):593-600.

[10] 李金海,邓硕.概念格与三支决策及其研究展望[J].西北大学学报:自然科学版,2017,47(3):321-329.

[11] Li J H,Mei C L,Lv Y J.Incomplete decision contexts:Approximate concept construction,rule acquisition and knowledge reduction[J].International Journal of Approximate Reasoning,2013,54(1):149-165.

[12] 苏婷,于杰.基于q近邻的不完备数据三支决策聚类方法[J].计算机科学与探索,2016,10(6):875-883.

[13] 刘盾,梁德翠.广义三支决策与狭义三支决策[J].计算机科学与探索,2017,11(3):502-510.

[14] 董静,孙乐,冯元勇,等.中文实体关系抽取中的特征选择研究[J].中文信息学报,2007,21(4):80-85.

[15] 黄鑫,朱巧明,钱龙华.基于特征组合的中文实体关系抽取[J].微电子学与计算机,2010,27(4):198-200.

[16] 朱艳辉,田海龙,刘璟,等.基于三支决策的新闻情感关键句识别方法[J].山西大学学报:自然科学版,2015,38(4):595-600.

[17] 田海龙,朱艳辉,梁韬,等.基于三支决策的中文微博观点句识别研究[J].山东大学学报,2014,49(8):58-65.

[18] 刘克彬,李芳,刘磊,等.基于核函数中文关系自动抽取系统的实现[J].计算机研究与发展,2007,44(8):1406-1411.

[19] ACE2005.The Automatic Content Extraction(ACE) projects[EB/OL].(2007).http://www.ldc.upenn.edu/Projects/ACE/.