ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 202－210
DOI:10.16337/j.1004-9037.2015.01.020

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于多分类 SVM-KNN 的实体关系抽取方法

刘绍毓　周　杰　李弼程　席耀一　唐浩浩

(解放军信息工程大学信息系统工程学院,郑州,450001)

摘　要: 实体关系抽取是信息抽取领域的重要研究课题之一。传统的实体关系抽取研究注重于从实体对出现的上下文中提取词法和语义等特征,然后利用分类器(如 SVM)进行实体关系抽取,但该类方法忽略了分类器对实体抽取性能的影响。针对 SVM 分类器对超平面附近样本分类正确率低的问题,本文设计了一种基于双投票机制的 SVM 模糊样本选择方法。在此基础上,对确定区域样本直接使用 SVM 分类器进行分类,并利用 KNN 算法对模糊区域样本进行二次分类。在 SemEval-2010 评测任务提供的实体关系抽取数据上进行实验,实验结果表明该方法能较大提高实体关系抽取的性能。

关键词: 支持向量机;KNN;双投票;实体关系抽取

中图分类号: TP391　　**文献标志码:** A

Entity Relation Extraction Method Based on Multi-SVM-KNN Classifier

Liu Shaoyu, Zhou Jie, Li Bicheng, Xi Yaoyi, Tang Haohao

(Information System Engineering College, PLA Information Engineering University, Zhengzhou, 450001, China)

Abstract: Entity relation extraction is one of the most important researches in the field of information extraction. Previous researches focus on extracting various kinds of lexical or semantic features from the context where the related entities appeared, and one kind of classifiers (such as SVM) is used to extract the entity relation, but this kind of methods ignore the impact of the classifier performance on the entity relation extraction. Since SVM classifier has low accuracy for the test samples near the hyperplane, a method based on double-vote mechanism is designed for determining the fuzzy SVM samples. In the method, SVM classifier is used to classify the non-fuzzy samples directly; then, *k*-nearest neighbors (KNN) algorithm is applied to classify the fuzzy ones. The experiment on the data provided by SemEval-2010 evaluation task shows that the method can improve the performance of the entity relation extraction.

Key words: SVM; KNN; double-vote; entity relation extraction

引　　言

随着计算机的快速普及和互联网的迅猛发展,电子信息的爆炸式增长给人们带来了巨大的挑战。人们迫切需要借助一些自动化的工具从海量数据中迅速找到自己真正需要的信息,信息抽取技术研究

基金项目: 国家高计划研究发展计划("八六三"计划)(2011AA7032030D)资助项目;全军军事研究生课题(军事学 YJS1062)资助项目。

收稿日期: 2013-12-03;**修订日期:** 2014-03-25

应运而生。信息抽取包括命名实体识别、实体关系抽取等多方面研究内容。实体关系抽取作为信息抽取研究领域的重要课题,其目的是要确定文本中已标注好的实体对之间的关系。例如,输入一个带有标记实体的句子,如"＜e1＞People＜/e1＞ have been moving back into ＜e2＞downtown＜/e2＞.",系统能自动识别实体"people"和"downtown"的关系类型是"Entity-Destination"。

实体关系抽取对信息检索、本体学习、语义网络标注、篇章理解都有重要的研究意义。作为一项基础性研究,它在文本分类、机器翻译、自动文摘、自动问答系统等方面都具有非常重要的应用[1]。例如,提取新闻报道中事件涉及的人物、时间、地点、原因、结果等信息,提取人物简历中人名、性别、学历等信息,并以结构化的形式存储到数据库中,能使查询更简便、更准确。

实体关系抽取的研究由来已久,1998 年消息理解会议(Message understanding conference, MUC)首次引入了实体关系抽取任务,后来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),其中的一项评测任务就是实体关系识别。传统的实体关系抽取大多局限于命名实体(包括人名、地名、组织机构名等)之间少数几种类型的关系,如"雇佣关系"、"地理位置关系"等,并将其转化为分类问题,将待识别样本分类到已定义的关系类型。SemEval-2010 评测任务 8 提出了一项实体关系抽取新任务——普通名词实体关系的多分类,引发了实体关系抽取研究的新高潮[2]。

实体关系抽取按其是否需要训练数据可分为半监督学习方法、无监督学习方法和监督学习方法[3]。半监督学习方法通常采用种子模式自扩展的方法,通过统计学习的方法来自动获取关系新模式[4],或者通过不断迭代产生可信度高的实例来扩充训练数据[5],但是它对初始关系种子的质量要求较高,领域迁移时需要重新编写规则或构建高质量的关系种子。无监督学习方法[6,7]的研究才刚刚起步,其主要思想是根据实体对的特征信息对实体对进行聚类,然后选择合适的能描述各类关系的特征;该方法虽然不需要事先定义实体关系类型,具有领域无关性,但目前仍缺乏标准的评测语料和统一的评价标准。监督学习方法主要有基于特征的方法[8]、基于核函数的方法[9,10]和基于知识库的方法[11]。其中,基于特征的方法首先抽取与样本相关的特征并形成特征向量,然后通过机器学习的方法来训练关系抽取模型。基于核函数的方法首先获得两实体所在句子的结构特征,形成结构树并计算它们之间的相似度,然后训练支持核函数的分类器进行关系抽取。基于知识库的方法通过借助外部的知识库(如维基百科、DBPedia, YAGO, OpenCyc, FreeBase 或者其他领域知识库)来进行实体关系抽取。但是,基于核函数的方法训练和预测的速度相对较慢,对大规模数据具有局限性;基于知识库的方法对知识库的质量和完备性提出了很高的要求,虽然领域性强的知识库能保证其有较高的准确率,但却无法避免其领域移植性差的问题。

1 基于改进的多分类 SVM-KNN 方法的实体关系抽取

传统的实体关系抽取研究在使用 SVM 分类器进行多分类处理时,其涉及的类别数较少,针对的也是相对简单的命名实体之间的关系。然而当实体关系类型数增加且处理对象是普通名词之间的实体关系时,各种关系类别的最佳边界往往难以确定,存在样本交叠问题。很多研究者都发现,尽管 SVM 分类器具有较强的抗噪声能力和较高的分类正确率,但其对分布在超平面附近区域的样本分类效果较差。林琛等[12]提出了信息粒度的概念,对处于类交叠区域外的样本直接利用 SVM 分类器进行分类,对处于类交叠区域内的样本使用 KNN 分类器进行分类,取得了较好的文本分类效果,但其"类交叠区域"仅针对简单的二分类问题。李丽双等[13]利用修正的 SVM-KNN 组合算法进行专有名词的自动抽取,但该方法用于二次分类的模糊区域样本仅仅针对三类中预先设定的两类,如果分类类别数增加,这种"交叠"或"模糊"区域往往很难确定其最佳边界。

本文针对实体关系抽取中普通名词实体关系多分类模糊样本难以界定的问题,设计了一种基于双投票机制的 SVM 模糊样本选择方法,在利用 SVM 分类器对待测样本进行双投票后,将测试样本集分为两部分,对确定区域的样本直接使用 SVM 分类器进行分类,并使用 KNN 分类器对模糊样本进行二

次分类。本文实体关系抽取流程图如图 1 所示。

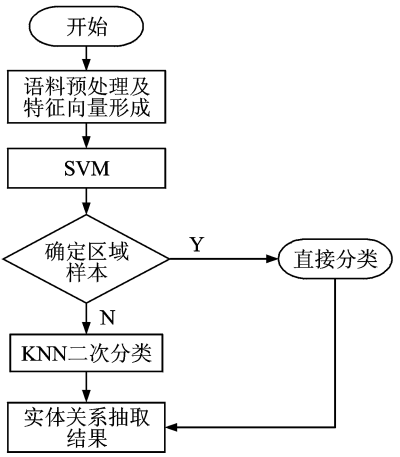

图 1 实体关系抽取示意图

Fig. 1 Architecture of entity relation extraction system

1.1 基于双投票的 SVM 模糊待测样本选择方法

在利用 SVM 分类器进行二分类时，可以根据待测样本到超平面的距离与最大间隔比值的大小来确定待测样本的模糊性。对多分类而言，SVM 分类器支持两种类型的分类方法：一种是 one-against-rest 方法，基本思想是对 $N(N\geqslant 3)$ 类的训练样本，把其中某类和其余各类样本的总和分别作为正、负例来训练分类器，但是这种方法往往会造成数据偏置的问题，从而导致泛化误差无界。另一种是 one-against-one 方法，即对 $N(N\geqslant 3)$ 类的训练样本，每两类样本训练一个分类器；在测试时计算待测样本的所有子分类器判决函数值，根据其正负采用投票的方法来确定最终类别[14]。对单个子分类器，其判决函数为

$$g(x)=\sum_i y_i\alpha_i K(x,\boldsymbol{x}_i)-b \tag{1}$$

式中：$\boldsymbol{x}_i$ 表示此子分类器的支持向量，b 为判决函数的偏置项，$K(\cdot)$ 表示核函数，$g(x)=0$ 表示超平面 H。其判决规则为：若 $g(x)>0$，待测样本被判为正类的投票数加 1；若 $g(x)<0$，待测样本被判为负类的投票数加 1。然后取最大投票数所对应的类别作为该待测样本的最终类别。针对 N 分类的某待测样本，假设利用 one-against-one 多分类方法对其进行投票的结果为：$\mathrm{Vote}=(v_1,v_2,v_3,\cdots,v_N)$，实验结果表明，$N$ 越大，投票数区分度越低。首先，最大投票数并非是该测试样本处于其对应类别确定区域（如图 2 所示）的投票数；其次，模糊区域不乏少数正类（负类）测试样本越过超平面而让其属于负类（正类）的投票增加的情况（如图 2 所示）。如果这时取 v_1 所对应的类别作为该待测样本的分类结果，可能会造成误判。

实际上，待测样本到超平面的距离 $D=|g(x)|/\|w\|$，半最大间隔为 $d=1/\|\boldsymbol{w}\|$（$\boldsymbol{w}$ 为支持向量），那么确定区域（如图 2 所示）样本满足 $|g(x)|\geqslant 1$，模糊区域（如图 2 虚线内区域所示）样本满足 $|g(x)|<1$。受此启发，设计一种新的投票方式：若 $g(x)\geqslant 1$，待测样本被判为正类的投票数加 1；若 $g(x)<-1$，待测样本被判为负类的投票数加 1；否则不投票。

SVM 的 one-against-one 多分类方式对待测样本的投票是一种粗略投票，但如果仅仅采用 $|g(x)|\geqslant 1$ 的标准来进行投票，也会因为条件过于严格而导致那些处于模糊区域的正、负例样本无法获得投票。综合这两种投票方式，本文设计了一种基于双投票机制的 SVM 模糊样本选择方法。

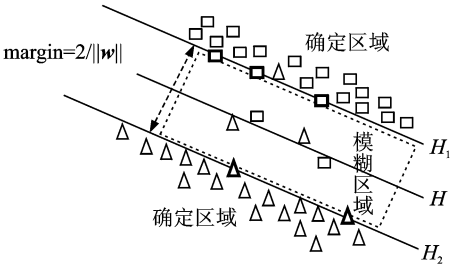

图 2 SVM 分类样本分布示意图

Fig. 2 Distribution of SVM classification samples

在采用 one-against-one 方式对类样本进行多分类时,设对某待测样本,根据判决函数 $g(x)$的正负进行投票并从高到低排列的结果序列是 $\mathrm{Vote}_1=(v_{11},v_{12},v_{13},\cdots,v_{1N})$,其对应的类别结果序列是 $\mathrm{ClassLabel}_1=(c_{11},c_{12},c_{13},\cdots,c_{1N})$,根据$|g(x)|\geqslant 1$ 的标准进行投票并从高到低排列的结果序列是 $\mathrm{Vote}_2=(v_{21},v_{22},v_{23},\cdots,v_{2N})$,其对应的类别结果序列是 $\mathrm{ClassLabel}_2=(c_{21},c_{22},c_{23},\cdots,c_{2N})$。对每次投票,可以比较合理地假设该待测样本的最终类别只在最大的两投票数对应的类别中出现,即可能在 $c_{11},c_{12},c_{21},c_{22}$所对应的类别中出现,若第二次投票差值 $\Delta=v_{21}-v_{22}$,第二次投票差值 Δ 足够大,说明最大投票数对应的类别严格"接纳"该待测样本的程度远远大于其他类别,这时待测样本属于确定样本集;若第二次投票差值 Δ 较大时,如果两次投票的最大投票数对应的类别相同,由两次投票的一致性可认为该待测样本属于确定样本集;其他情况下待测样本由于投票的区分度低而具有模糊性。现制定模糊样本选择规则如下:

设 T 是确定样本集,F 是模糊样本集,α 和 β 是整数阈值因子。

(1)当 $\Delta\geqslant\alpha$ 时,若 $c_{11}=c_{21}$,将样本加入 T;若 $c_{11}\neq c_{21}$,将样本加入 F,并标记为模糊样本子集 $F(c_{11},c_{21})$。

(2)当 $\beta\leqslant\Delta\leqslant\alpha$ 时,若 $c_{11}=c_{21}$ 且 $c_{12}=c_{22}$,将样本加入 T;若 $c_{11}=c_{21}$ 且 $c_{12}\neq c_{22}$,或 $c_{12}\neq c_{21}$ 且 $c_{11}=c_{22}$,将样本加入 F,并标记为模糊样本子集 $F(c_{21},c_{22},c_{12})$。若 $c_{11}\neq c_{21}$ 且 $c_{12}=c_{22}$,或 $c_{12}=c_{21}$ 且 $c_{11}\neq c_{22}$,将样本加入 F,并标记为模糊样本子集 $F(c_{21},c_{22},c_{11})$。若 $c_{11},c_{12},c_{21},c_{22}$ 互不相等,将样本加入 F,并标记为模糊样本子集 $F(c_{21},c_{22},c_{11},c_{12})$。

(3)当 $\Delta\leqslant\beta$ 时,将样本加入 F,其中若 $c_{11}=c_{21}$ 且 $c_{12}=c_{22}$,将样本标记为模糊样本子集 $F(c_{11},c_{12})$,其余与第二种情况相同。

1.2 基于多分类的 SVM-KNN 分类算法

KNN 分类算法是一种简单易行的无参数分类方法,在使用该方法进行分类时选出距待测样本 K 个最近的训练样本,将多数样本对应的类别确定为该待测样本的类别。本文利用 KNN 算法对前述双投票方法选择出来的待测样本进行二次分类。SVM-KNN 算法能有效改善 SVM 分类器的分类效果,其基本思想是:将测试样本集输入到 SVM 分类器并采用前述的基于双投票的模糊样本选择方法对其进行处理,形成确定样本集 T 和模糊样本集 F(包括多个模糊样本子集 F^i)。对确定样本集 T 中的样本,直接采用 SVM 分类器进行分类;对模糊样本子集 F 中的样本,采用 KNN 分类器进行分类,这时将 SVM 分类器训练得到的支持向量集作为 KNN 分类器的训练集。

输入:训练集 E、测试集 C,类别标记集 L,确定集决定规则 T_{law} 与模糊样本选择规则 F^i_{law}(由第二次投票差值 Δ 和整数阈值因子 α 和 β 决定),KNN 训练取值集合 K。

输出:各测试样本的预测类别

训练阶段：

01：将训练集 E 形成的特征向量输入到 SVM 分类器中进行处理，得到分类模型的支持向量集 $SvSet$。

02：计算 $SvSet$ 中任意两支持向量的距离，形成支持向量集类别—距离矩阵 $\boldsymbol{P}$。

测试阶段：

01：for $c_j \in C$

02：if c_j subject to T_{law}，直接利用 SVM 分类器进行分类。

03：else calculate $FL(c_j)$ and let $c_j \rightarrow F^i(FL(c_j))$（其中 $FL(c_j)$ 为样本 c_j 所对应的模糊类标记集，F^i 为对应的模糊样本子集）

04：end for

05：let all $F^i \rightarrow F$

利用 KNN 分类器对 F 进行二次分类

06：for each $F^i \subseteq F$

07：根据支持向量集类别-距离矩阵 $\boldsymbol{P}$ 寻找最优 KNN 分类器（由最优 $K_0 \in K$ 决定）

08：对每个 $c \in F^i \subseteq F$ 利用训练好的 KNN 分类器对其进行分类。

09：end for

在利用 KNN 分类器进行分类时，常用的距离计算方式有两种，一种是欧式距离，另一种是曼哈顿距离。设两特征向量 $\boldsymbol{X}=(x_1, x_2, x_2, \cdots, x_l)$ 和 $\boldsymbol{Y}=(y_1, y_2, y_2, \cdots, y_l)$，它们的距离计算公式如下

欧式距离 $$D_{\text{E}}(\boldsymbol{X},\boldsymbol{Y})=\sqrt{\sum_{i=1}^{l}(x_i-y_i)^2} \tag{2}$$

曼哈顿距离 $$D_{\text{M}}(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{l}\| x_i-y_i \| \tag{3}$$

2 实验及性能比较

2.1 实验准备及实验方案

实验语料选择 SemEval-2010 评测任务中的实体关系抽取语料集，该语料集将实体关系标注为 10 种类型，分别是：Component-Whole，Instrument-Agency，Member-Collection，Cause-Effect，Entity-Destination，Content-Container，Message-Topic，Product-Producer，Entity-Origin，Other。其中“Other”类是根据标注规则不属于前 9 类实体关系任何一种的总和，而前九类实体关系又有正反顺序之分，（如 Component-whole 简写为 CW，CW1 表示关系句子实例中的第一个实体为 Component，第二个实体为 Whole，而 CW2 与之相反）所以共标注为 19 种类型。其训练语料包含已标注好实体和关系类型的 8 000 条句子，本文实验将其前 10％句子作为测试集，其余的 90％句子作为训练集。

语料预处理包括词性标注、词干提取、句法分析、谓词提取和语义角色标注等。本文采用的特征有：实体及上下文特征（包括实体及其前后的词、词的词干和词性）、句子动词词根特征、实体距离特征、实体扩展特征、语义角色特征和实体间词语特征，其中语义角色特征又包括谓词特征、语义角色对特征、语义角色对-谓词特征。实体及其上下文词的词干和词性抽取分别采用波特词干抽取法和兰卡斯特大学提供的词性抽取算法。实体扩展特征由 Cycorp 提供，并采用 CCG Curator 进行语义角色标注[15]。以句子“The metal <e1>ball</e1> makes a ding ding ding <e2>noise</e2> when it swings back and hits the metal body of the lamp.”为例，其实体及上下文特征如表 1 所示。

其动词特征有 make，swing，hit，实体距离特征是 5。实体“ball”的扩展特征有 CompositePhysicalAndMentalEvent，SocialOccurrence，CulturalThing 等，实体“noise”的扩展特征有 Container-Underspecified，Path-Underspecified，Path-Generic 等。语义角色特征包括谓词特征 make，语义角色对特征（A0，

A1)及其组合特征(A0,make,A1)。实体间词语特征有 make,a,ding,ding,ding。

表 1 实体及上下文特征表示(n-gram=2)

Table 1 Features of entities and their context

位置	实体 1(−2)	实体 1(−1)	实体 1	实体 1(+1)	实体 1(+2)	实体 2(−2)	实体 2(−1)	实体 2	实体 2(+1)	实体 2(+2)
词	The	metal	ball	makes	a	ding	ding	noise	when	it
词根	the	metal	ball	make	a	ding	ding	noise	when	it
词性	ATO	NN1	NN1	VVZ	ATO	AJO	AJO	NN1	CJS	PNP

在对实验语料进行预处理后,构造特征向量;训练 SVM 分类模型,获得每类训练样本的支持向量;计算得到支持向量集的类别-距离矩阵,形成 SVM-KNN 分类框架;将测试样本输入到 SVM-KNN 分类器,判断测试实例的关系类型。

2.2 实验结果及其分析

经实验发现,三类或四类模糊的待测样本极少,本文实验的二次分类只对二类模糊进行处理,并采用常用的准确率、召回率、$F1$ 值对其进行评价[16]。通过迭代实验,设置最佳整数阈值因子 $\alpha=3$,$\beta=1$。实验结果与文献[17]的方法(目前在该语料集上的最佳结果)进行对比分析,该方法在使用词汇特征、实体扩展特征和动词等特征的基础上,将谓词特征和语义角色对特征进行融合,并引入实体间词语特征,利用 SVM 分类器进行实验。本文以其研究为基础,并在 KNN 分类器训练和测试时采用欧式方法和曼哈顿方法计算样本距离。

从表 2 所示的实验结果可以看出,本文方法的准确率和召回率有了较大提高,$F1$ 值分别提高了 9.5%和 9.3%,显著提高了实体关系抽取的性能。使用支持向量集训练的 KNN 分类器比直接使用训练集训练的 KNN 分类器具有更高的实体关系抽取准确率,这是因为从训练集中剔除了那些对各个关系类别代表性不强的样本。这样做虽然提高了分类器的精度,也使各关系类别接纳测试样本的条件更苛刻,从而导致召回率下降,但 F 值相差不大,综合考虑计算量,应优先选择支持向量集作为 KNN 分类器训练集。

表 2 本文方法与文献[17]方法的实体关系抽取结果比较

Table 2 Comparison of the results of entity relation extraction between our method and the method in Ref.[17]

实验标号	方法	KNN 训练语料	准确率/%	召回率/%	$F1$ 值/%
1	文献[17]的方法		78.1	85.4	81.6
2	SVM−KNN (动态 K+欧式距离)	训练集	90.4	91.7	91.1
3	SVM−KNN (动态 K+欧式距离)	支持向量集	90.7	91.1	90.9

在二次分类时,对 KNN 分类器分别采用欧式方法和曼哈顿方法计算距离,并比较 SVM-KNN 分类算法在 K 取值 1,3,5,7,9,11 和动态 K 的抽取性能。从图 3 中可以看出,当 K 取不同值时,基于 SVM-KNN 的方法都有较高抽取性能,充分说明本文方法能较大程度提高实体关系抽取性能。K 值固定时,采用曼哈顿距离计算方法的分类器比采用欧式距离计算方法的分类器具有较高的准确率和较低召回率。但由于曼哈顿距离计算复杂度低,所以在 $F1$ 值相差不大的情况下,应优先考虑采用曼哈顿距离计算方式训练的 KNN 分类器,此时 $F1$ 值大致随着 K 的增大先增后降,$K=7$ 时,$F1$ 值最大;此时动态 K 的 SVM-KNN 方法抽取准确率较高,召回率最低,这是因为训练过程中产生了过拟合现象,即使这样,

效果仍然较好。表 3 为采用动态 K 的 SVM-KNN 分类器、利用曼哈顿方法计算样本距离时的实体关系抽取结果。

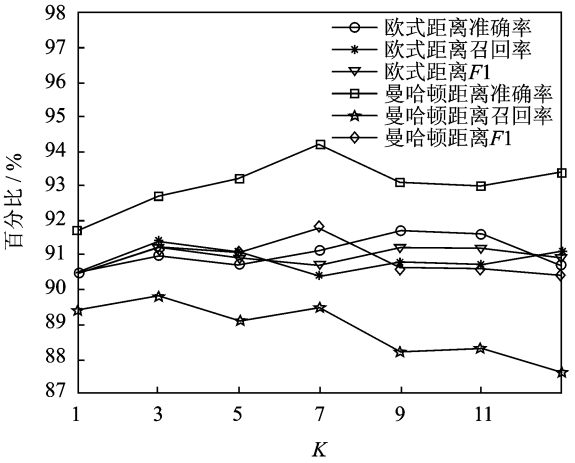

图 3 不同距离计算方式和 K 值下的实体关系抽取结果

Fig. 3 Results of entity relation extraction by using different distance calculations and K values

表 3 基于改进的 SVM-KNN 方法的实体关系抽取结果分类表示

Table 3 Results of entity relation extraction based on the method of improved SVM-KNN classification

实体关系类型		准确率/%	召回率/%	$F1$ 值/%
Cause-Effect	CE1	100	84.6	91.7
	CE2	95.8	98.6	97.1
Component-Whole	CW1	97.3	90.0	93.5
	CW2	93.9	73.8	87.7
Content-Container	CC1	100	90.9	95.2
	CC2	89.1	90.5	89.8
Entity-Destination	ED1	98.8	82.0	89.6
	ED2	—	—	—
Entity-Origin	EO1	95.7	83.0	88.9
	EO2	88.9	80.0	84.2
Instrument-Agency	IA1	87.8	100	93.3
	IA2	93.2	83.7	88.2
Member-Collection	MC1	75.0	85.7	80.0
	MC2	89.2	100	94.3
Message-Topic	MT1	89.1	90.5	89.8
	MT2	100	53.8	70.0
Product-Producer	PP1	85.2	85.2	85.2
	PP2	93.1	81.8	87.1
Other		60.6	81.3	69.4
Overall		93.4	87.6	90.4

一般来说，除了“Other”类外，其他各类都获得了较高的准确率，甚至有些类别的准确率达到了 100%。导致这一结果，一方面，本文提出的模糊样本选择方法能有效缩小样本的类别模糊空间；另一方

面,SVM-KNN 分类器能够明确划定各类的边界,具有较高的分类精度。高准确率虽然导致有些关系类别的召回率受到影响,但从整体上看,除了少数类别实体关系抽取性能稍差外,其他各类都取得了较好效果。另外,相比于文献[17]的方法,本文方法"Other"类的召回率得到了显著提高,达到了 81.3%,这表明"Other"类被误判为其他类的几率大大减少,从而使整体性能获得了较大提升。

3 结束语

传统的机器学习方法一般通过选择不同的表示模型、特征、分类器等来提高分类性能。SVM 分类器作为自然语言处理常用的分类器,具有较强的抗噪声能力和较好的分类性能,但是在使用 one-against-rest 的方式进行多分类时,往往处于模糊区域(两分类面内部)的样本会因为基于正负规则的判决方式而产生投票误增现象,最终导致样本的各类投票数差距不明显而造成错分。为了克服 SVM 分类器的这一弱点,本文设计了一种基于双投票机制的 SVM 模糊样本选择方法,该方法能有效区分确定区域样本和模糊区域样本,对确定区域样本直接使用 SVM 分类器进行分类,对模糊区域样本使用 KNN 分类器进行分类。实验结果表明,该方法降低了模糊待测样本识别的错误率,有效提高了实体关系的抽取性能。

参考文献:

[1] 张传岩,洪晓光,彭朝晖,等. 基于 SVM 和扩展条件随机场的 Web 实体活动抽取[J]. 软件学报,2012,23(10):2612-2627.
Zhang Chuanyan, Hong Xiaoguang, Peng Zhaohui, et al. Web entity activities based on SVM and extended conditional random fields[J]. Journal of Software, 2012,23(10):2612-2627.

[2] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C]//Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. [S. l.]: Association for Computational Linguistics, 2009:94-99.

[3] 宁海燕. 实体关系自动抽取技术的比较研究[D]. 哈尔滨:哈尔滨工业大学,2010.
Ning Haiyan. Comparative study of automatic entity relation extraction[D]. Harbin: School of Computer Science and Technology of Harbin Institute of Technology, 2010.

[4] 邓擘,郑彦宁,傅继彬. 汉语实体关系模式的自动获取研究[J]. 计算机科学,2010,37(2):183-185.
Deng Bo, Zheng Yanning, Fu Jibin. Study of obtaining Chinese entity relation pattern automatically[J]. Computer Science, 2010,37(2):183-185.

[5] 郭剑毅,雷春雅,余正涛,等. 基于信息熵的半监督领域实体关系抽取研究[J]. 山东大学学报:工学版,2011,41(4):7-12.
Guo Jianyi, Lei Chunya, Yu Zhengtao, et al. A semi-supervised learning method based on information entropy to extract the domain entity relation[J]. Journal of Shangdong University: Engineering Science, 2011,41(4):7-12.

[6] Rink B, Harabagiu S. A generative model for unsupervised discovery of relations and argument classes from clinical texts [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. [S. l.]: Association for Computational Linguistics, 2011:519-528.

[7] 黄晨,钱龙华,周国栋,等. 基于卷积树核的无指导中文实体关系抽取研究[J]. 中文信息学报,2010,24(4):11-17.
Huang Chen, Qian Longhua, Zhou Guodong, et al. Research on unsupervised Chinese entity relation extraction based on convolution tree kernel[J]. Journal of Chinese Information Processing, 2010,24(4):11-17.

[8] Tratz S, Hovy E. ISI: Automatic classification of relations between nominals using a maximum entropy classifier[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. [S. l.]: Association for Computational Linguistics, 2010:222-225.

[9] 李丽双,党延忠,张婧,等. 基于组合核的中文实体关系抽取研究[J]. 情报学报,2012,31(7):702-708.
Li Lishuang, Dang Yanzhong, Zhang Jing, et al. Chinese relation extraction based on ensemble kernel[J]. Journal of the China Society for Scientific and Technical Information, 2012,31(7):702-708.

[10] Choi S P, Lee S, Jung H, et al. An intensive case study on kernel-based relation extraction[J]. Multimedia Tools and Applications, 2013:1-27.

[11] 张苇如,孙乐,韩先培. 基于维基百科和模式聚类的实体关系抽取方法[J]. 中文信息学报,2012,26(2):75-81.

Zhang Weiru, Sun Le, Han Xianpei. A entity relation method based on wikipedia and pattern clustering[J]. Journal of Chinese Information Processing, 2012,26(2):75-81.

[12] 林琛,李弼程,周杰. 基于信息粒度的交叠类文本分类方法[J]. 情报学报,2011,30(4):339-346.
Lin Chen, Li Bicheng, Zhou Jie. A text categorization method for overlapping classes based on information granularity[J]. Journal of the China Society for Scientific and Technical Information, 2011,30(4):339-346.

[13] 李丽双,党延忠,李丹. 基于修正 SVM-KNN 组合算法的汉语专有名词自动抽取[J]. 情报学报,2011,30(6):610-617.
Li Lishuang, Dang Yanzhong, Li Dan. Automatic extraction on Chinese proper names based on a modified SVM-KNN classifier[J]. Journal of the China Society for Scientific and Technical Information, 2011,30(6):610-617.

[14] 闫志刚,杜培军. 多类支持向量机推广性能分析[J]. 数据采集与处理,2009,24(4):469-475.
Yan Zhigang, Du Peijun. Generalization performance analysis of M-SVMs[J]. Journal of Data Acquisition and Processing, 2009,24(4):469-475.

[15] Punyakanok V, Roth D, Yih W. The importance of syntactic parsing and inference in semantic role labeling[J]. Computational Linguistics, 2008,34(2):257-287.

[16] 李天颖,刘璘,赵德旺,等. 一种基于依存文法的需求文本策略依赖关系抽取方法[J]. 计算机学报,2013,36(1):54-62.
Li Tianying, Liu Lin, Zhao Dewang. Eliciting relations from requirements text based on dependency analysis[J]. Chinese Journal of Computers, 2013,36(1):54-62.

[17] 毛小丽,何中市,邢欣来,等. 基于语义角色的实体关系抽取[J]. 计算机工程,2011,37(17):143-145.
Mao Xiaoli, He Zhongshi, Xing Xinlai, et al. Entity relation extraction based on semantic role[J]. Computer Science, 2011, 37(17):143-145.

作者简介:**刘绍毓**(1987-),男,硕士研究生,研究方向:实体关系抽取,E-mail:tyughvbn88@163.com;**周杰**(1984-),男,博士研究生,研究方向:人名消歧;**李弼程**(1970-),男,教授,博士生导师,研究方向:语音信号处理与智能信息处理;**席耀一**(1987-),男,博士研究生,研究方向:信息抽取;**唐浩浩**(1990-),男,硕士研究生,研究方向:情感分析。