# 基于SVM+Sigmoid的汉语组块识别*)

谭咏梅 姚天顺 陈 晴 李 珩 朱靖波

(东北大学信息学院软件所自然语言处理实验室 沈阳110004)

**摘 要** 本文提出用SVM+Sigmoid来进行汉语组块识别的方法。SVMs具有不需要进行认真选取特征的优点,并且在具有高维特征空间的输入数据上也能够具有高的泛化性能,通过核函数的原则,SVMs能够在独立于训练数据维数的小计算范围内进行训练。Sigmoid函数使用一个参数模型来直接拟合后验概率,从而将SVMs的输出映射成一个后验概率,使一个分类器在做全局决策的一个局部决策时,考虑到全面分类,从而决策更具有合理性。实验结果表明该方法较单纯的SVMs方法具有好的效果。

**关键词** 支持向量机,Sigmoid函数,汉语组块,组块识别

## Support Vector Machines Plus Sigmoid Based Chinese Chunk Recognition

TAN Yong-Mei YAO Tian-Shun CHEN Qing LI Heng ZHU Jing-Bo

(Natural Language Processing Lab ,Northeastern University,Shenyang 110004)

**Abstract** The paper presents a method of Chinese Chunk Recognition based on Support Vector Machines(SVMs) plus Sigmoid. Conventional recognition techniques based on Machine Learning have difficulty in selecting useful features as well as finding appropriate combination of selected features. On the other hand,it is well known that SVMs achieve high generalization of very high dimensional feature space. Furthermore,by introducing the Kernel principle, SVMs can carry out the training in high-dimensional space with smaller computational cost independent of their dimensionality. Sigmoid function is a method of extracting probabilities (class/input)for SVMs outputs,which is helpful to classification post-processing. Sigmoid allows decisions that can use a utility mode,also is needed when a classifier is making a small part of an overall decision,and the classification outputs must be combined for the overall decision. The experiments produce promising results.

**Keywords** Support vector machines(SVM),Sigmoid function,Chinese chunk,Chunk recognition

## 1 介绍

组块识别是自然语言处理中的一个新兴的研究课题,它的任务往往是在不需要深层次语言知识的前提下,识别句子中的某些特定成分,即从一系列的词串中识别出具有句法关系的某些子串。

组块识别作为一种预处理手段,可以大大降低进行短语划分和短语分析处理的复杂性,为进一步对句子的深层次分析提供了基础,使得句法分析任务在某种程度上得到简化,同时对机器翻译、信息提取、信息检索、专有名词识别等都具有非常重要的意义。

组块处理的方法主要有下面的两种,一种是基于规则的方法;另一种是统计的方法。现在人们更多地使用机器学习的技术来进行文本组块处理,因为机器学习能够避免单调乏味的手工工作,并且有助于性能的提高。

本文定义了7种汉语组块类型:基本副词短语(BDP),基本形容词短语(BAP),基本数量短语(BMP),基本时间短语(BTP),基本处所短语(BNS),基本名词短语(BNP),基本动词短语(BVP)。虽然文[1]定义了9种汉语组块类型,但是由于区别词(bp)从词汇层面说,它固有的句法功能是只能作定语或跟助词"的"构成"的"字结构[2],因此本文没有定义区别词短语,并且因为准数词短语(mbar)与基本数词短语(BMP)的句法结构功能都是相同的,所以就将准数词短语(mbar)归入基本数词短语(BMP)。

在英文组块(chunk)识别方面,文[3]使用支持向量机的机器学习方法进行了chunk的识别,在CoNLL-2000的共享任务组块处理中取得了最好的性能。

在汉语组块识别方面,国内已有人做过一些有益的工作。例如张琪等[1]利用短语内部结构和词汇信息对预测中出现的边界歧义和短语类型歧义进行了排岐处理。周强等[4]介绍了汉语句子的组块分析体系,通过引入词界块和成分组概念,将成分识别问题从完整的句法分析任务中分离出来。赵军等[5]从语言学的系统角度定义了汉语基本名词短语,提出了将汉语基本名词短语的结构模板和其上下文环境特征结合的汉语基本名词短语识别模型。以上的研究都取得了成功,但是他们所使用的都是非公开的语料,并且对汉语组块的定义也不完全相同,所以无法直接比较模型性能的优劣。

SVMs是一个学习分类器的有效方法,这种方法已经成

*)基金项目:国家自然科学基金资助项目(60083006);国家重点基础研究发展规划973资助项目(G19980305011);国家自然科学基金和微软亚洲研究院基金资助(60203019)。**谭咏梅** 博士研究生,主要研究方向为机器翻译、自然语言处理;**姚天顺** 教授,博士生导师,主要研究方向为机器翻译、自然语言处理;**陈 晴** 硕士研究生,主要研究方向为机器翻译、自然语言处理;**李 珩** 博士研究生,主要研究方向为机器翻译、自然语言处理;**朱靖波** 副教授,博士,主要研究方向为自然语言处理。

功地运用在了自然语言处理的任务，如英语组块识别[3,6]，英语词性标记[7]，英语专名识别[8]和文本分类[9]等方面，并在这些领域都取得了不错的效果，但是在汉语组块分析方面却没有相关的报道，所以本文使用了 SVMs 来进行汉语组块的识别。

考虑到出现两个或多个分类器对于一个未知类别的样本给出两个或多个类别，或没有一个分类器对于一个未知类别的样本给出类别标识的情况下，在 SVMs 的基础上训练了 Sigmoid 函数参数，将 SVMs 的输出转化为一个校准的后验概率，从而一个分类器在做全局决策的一个局部决策时，考虑了全面分类，使得所作出的决策更合理。

本文第2节定义了7种汉语组块类型，第3节介绍了支持向量机，第4节描述了基于支持向量机概率的汉语组块识别，最后是结束语。

## 2 汉语组块类型

Chunk 是 Abney 在1991年提出的[10]，其目的是用于句法分析，他建议开发一个基于 Chunk 的句法分析器，首先将一个句子分成若干个 Chunk，其中每一个 Chunk 是具有单一语义核心和严格非递归句法结构的单词和连续词串，句子中的每一个单词都将划到某个组块中。

CONLL2000的共享任务组块处理中，采用 Abney 的组块描述框架和宾州树库的华尔街日报(WSJ)部分，开发了一个具有一定规模的英语 chunk 库，为基于统计的各种不同英语句法分析方法提供了统一的评价标准[11]。

与 Abney 定义的 Chunk 类似，我们定义的中文组块具有单一语义核心，并具互相不嵌套的特点，即句子中的每一个单词只能属于一个组块类型，并且每一种组块类型中都不含有其他类型的组块。考虑到汉语和英语是两种不同的语系，因此在汉语的情况下，定义了7种汉语组块类型：

1)基本副词短语(BDP)：一般是以副词为核心词的短语。

2)基本形容词短语(BAP)：指核心词为形容词的短语。

3)基本数量短语(BMP)：该类型中的量词核心指时间量词之外的量词。

4)基本时间短语(BTP)：指核心词是表示时间的量词、时间词，和部分表示时间的名词，以及这些结构的混合并列构成的短语。

5)基本处所短语(BNS)：表示的是基本短语中的表示地点、地域或者表示这些概念的名词结构。

6)基本名词短语(BNP)：是一类紧密结合的名词结构，由修饰词＋名词核心词构成。

7)基本动词短语(BVP)：包括：动趋搭配、动补搭配、形式动词加实意动词，动词被"不/得"分割等等现象。

本文使用下面的3种组块类型的边界标记来标识汉语组块：

1) B(begin)-XP XP∈{BDP, DAP, BMP, BNS, SNP, BVP} 表示该词是一个类型为 XP 的新组块的开始。

2) I(inside)-XP XP∈{BDP, DAP, BMP, BNS, SNP, BVP} 表示该词属于类型为 XP 的组块中。

3)O(outside)表示该词不属于任何类型的组块。

因此一种类型为 XP 的组块有两种边界标记 I-XP 和 B-XP，本文定义了7种组块类型，所以在实验中有14种边界标记，加上不属于任何组块类型的边界标记 O，一共15种。

使用上面定义的组块类型边界标记，可以将汉语组块识别看作是一个分类的问题：即，输入一个经过分词和词性标注的汉语句子，汉语组块识别器将每一个词标注上组块类型边界标记，输出结果为一序列的组块类型边界标记。下面给出一个具体的分析实例。

表1 汉语组块自动识别的例子

| 输入 | 小/h | 刘/nx | 现在/t | 正/d | 忙/vg | 着/ut | 开会/vg | 呢/y | 。/wj |
|---|---|---|---|---|---|---|---|---|---|
| 输出 | B-BNP | I-BNP | O | B-BVP | I-BVP | O | B-BVP | O | O |
| 组块类型 | B-BNP | | O | B-BVP | | O | BVP | O | |

## 3 支持向量机

### 3.1 支持向量机(SVMs)

SVMs 是 Vladimir Vapnik 提出的，最早可追溯到1979年，但是关于 SVMs 的第一篇详细文章却是在1995年。SVMs 是一个进行二值分类比较新的学习方法，其基本思想是寻找一个能够将 $d$ 维空间中的数据正确地分为两个类别的超平面。

SVMs 是在统计学习理论的理论框架下产生出来的，由于统计学习理论为人们研究小样本情况下机器学习问题提供了有力的理论基础，通过控制学习机器的容量来实现对推广能力的控制，在小样本问题中要遵循结构最小化(SRM)归纳原则，即针对经验风险和置信范围这两项最小化风险泛函[12]，SVMs 保持经验风险值固定而最小化置信范围，所以表现出了令人向往的优良特性。

SVMs 象其他的归纳学习方法一样，将输入作为一个训练实例的集合，寻找一个将它们映射到一个类别的分类函数。假定我们具有一个两个类别问题的训练数据集合：$(x_1, y_1)$，…，$(x_l, y_l)$，这里$(x_i \in R^d)$是第 $i$ 个样本在训练数据中的特征向量，$y_i \in \{-1, +1\}$是第 $i$ 个样本的类别标记，$y_i = +1$表示第 $i$ 个样本属于正例，$y_i = -1$表示第 $i$ 个样本属于反例，SVMs 通过训练构建一个能够对未知类别的样本进行正确类别预测的函数 $g(x)$。则指示函数 $f(x)$为

$$f(x)=sign(g(x))=\begin{cases}+1 & \text{if } g(x)>0\\ -1 & \text{otherwise}\end{cases}$$

如何构建这个正确的分类的函数 $g(x)$呢?通过在训练过程中寻找一个最优的超平面，也就是间隔最大的平面，间隔定义为这个平面和与这个平面最近的训练样本(支持向量)之间的距离。在 SVMs 中仅支持向量对构造最优超平面起作用。假设 $k$ 是训练样本中的支持向量的数目，则构建这个超平面的 VC 维(评价一个系统在未知数据上性能优良的指标)与 $k$ 成比例。

假设训练数据可被一个超平面：

$$(w\cdot x)-b=0 \quad w\in R^n, b\in R \tag{1}$$

没有错误地分开，则分割超平面可写为：

$$y_i[(w\cdot x_i)-b]\geqslant 1 \quad (i=1,\cdots,l) \tag{2}$$

样本点 $x_i$ 到分割超平面的距离可以表示为：

$$d(w,b;x_i)=\frac{|g(x_i)|}{\|w\|}=\frac{|wx_i+b|}{\|w\|}$$

则两个分割超平面之间的间隔表示为：

$$\underset{x_i,y_i=1}{d}(w,b;x_i)+\underset{x_i,y_i=-1}{d}(w,b;x_i)=\underset{x_i,y_i=1}{d}\frac{|w\cdot x_i+b|}{\|w\|}+\underset{x_i,y_i=-1}{d}\frac{|w\cdot x_i+b|}{\|w\|}=\frac{2}{\|w\|}$$

为了最大化这个间隔，必须使$\|w\|$最小，从而SVMs的训练问题转换为下面的优化问题。

Minimize $\frac{1}{2}\|w\|^2$

约束条件：$y_i[(w\cdot x_i)-b]\geqslant 1 \quad (i=1,\cdots,l)$

如果训练数据不能够被一个超平面没有错误的分开，则引入了核函数$k(x,x_i)$，将训练数据$x_i$映射到一个高维的特征空间，使得在高维空间中线性可分，于是在高维空间中来构造最优超平面。根据不同的核函数，可以实现输入空间中不同类型的非线性决策面的学习机器。如多项式核函数：$k(x,x_i)=[(x,x_i)+1]^d$，$d$是需要选取的参数，在实际应用中取值在1到10的范围内，高斯径向基核函数$k(x,x_i)=\exp(-\frac{\|x-x_i\|}{2\delta^2})$依赖于两个向量之间的距离$\|x-x_i\|$，这里$\delta$是需要调整的参数，两层神经网络核函数$k(x,x_i)=s[u(x\cdot x_i)+c]$，其中$s(u)$是sigmoid函数[13]。

### 3.2 多分类支持向量机(Multiclass SVMs)

SVMs构建了一个二值分类器，仅能够对两个类别进行分类，在多类别的情况下需要将SVMs扩展到多个类别的分类器。对于$n$个类别的分类问题，目前构造多分类器的方法有如下两种：

1. One-against-all：从$n$个类别中为每一个类别创建二值分类问题，也就是，对于每一个类别$i(1\leqslant i<n)$，二值分类问题使用给定的基于间隔的学习算法，从而构造$n$个SVMs二值分类器$c_i, i\in\{1,2\cdots,N\}$，这里标记为$y=i$的实例被认为是正例，所有其余标记的实例被认为是反例。

2. Pairwise：在任意类别$i$和类别$j(1\leqslant i,j\leqslant n, i\neq j)$之间构造一个SVMs二值分类器，从而生成了$n(n-1)/2$个SVMs二值分类器$c_i(1\leqslant i\in\{1,2\cdots,n(n-1)/2\})$，使用给定的二值学习算法来区别所有类别的每一对，这里标记为$y=i$的实例被认为是正例，标记为$y=j$的实例则被认为是反例，对于一个未知样本每一个分类器都有一个选票，其结果是具有选票最多的类别。

## 4 基于支持向量机概率输出的汉语组块识别

### 4.1 输入特征

SVMs不需要认真地进行特征选取来取得好的结果，因为它们能够在给定的特征集中自动进行选取。在特征的总维数非常大情况下也具有很好的泛化能力，并能够在具有多种特征组合的情况下进行训练。所以我们将出现在训练数据中不同位置的所有单词$w$、词性标记$p$和组块类型边界标记$t$作为特征，充分利用当前标记位置的上下文信息，将每一个样本$x$用13个特征$f$来表示：

$x=(w_{-2},p_{-2},t_{-2},w_{-1},p_{-1},t_{-1},w_0,p_0,w_{+1},p_{+1},w_{+2},p_{+2})$；

在本文中，假设分类过程是从左到右进行的，这个可以从下面的特征定义中看到。

$w_0$：表示当前位置的单词，$p_0$表示$w_0$的词性标记，$t_0$表示$w_0$的组块类型边界标记，是我们要进行分类的组块类型边界标记。

$w_{-i}$：表示从当前位置往前数第$i$个的单词，$p_{-i}$表示$w_{-i}$的词性标记，$t_{-i}$表示$w_{-i}$的组块类型边界标记。

$w_{+i}$：表示从当前位置往后数第$i$个的单词，$p_{+i}$表示$w_{+i}$的词性标记。

对于特征SVMs二值分类器仅接受数字化的值，为了满足这个限制，通过构建一个关于特征的倒排索引表InvTab，其中的每个记录为二元组$(f,\text{index}_w)$，其中index是特征$f$所在的特征列表中的位置。如$(w_{-2}=\text{美丽},1001)$，表示"$w_{-2}=$美丽"这个特征是特征列表中的第1001个元素。特征倒排索引表按特征以散列方式组织，能够实现快速查找，从而很方便地将每一个样本的特征用一系列的数字来表示。

### 4.2 多分类

设任意一个有$n$个词的汉语句子表示为：$W^T=w_1w_2\cdots w_n$，$w_i(1\leqslant i\leqslant n)$表示第$i$个单词，该汉语句子的词性序列为$p^T=p_1p_2\cdots p_n$，$p_i(1\leqslant i\leqslant n)$表示$w_i$的词性，该汉语句子的组块类型边界标记序列为$T^T=t_1,t_2\cdots t_n$，$t_i(1\leqslant i\leqslant n)$表示$W^T$所对应的组块类型边界标记。则汉语组块识别问题，可以看作是在给定词序列$W^T$和其所对应的词性序列$T^T$的情况下，将每一个$w_i(1\leqslant i\leqslant n)$所对应的组块类型边界标记$t_i(1\leqslant i\leqslant n)$在前面定义的15种组块类型边界标记中进行分类的过程。由于所使用的语料规模小，如果采用pairwise的多分类器扩展方法，则要生成$\frac{15\times(15-1)}{2}=105$个二值SVMs分类器，这样会使得每一份训练语料非常小，从而使得分类效果下降，因此采用one-against-all方法来进行实验。实验语料是实验室内部的500KB树库，其中90%作为训练语料，10%作为测试语料。

实验1：根据one-against-all的多分类器扩展方法，产生了15个二值SVMs分类器。每一个类别分类器$f_i(x)$被训练来将第$i$类的成员从其他类别的成员中区分开。使用给定的二值学习算法来区别所有类别中的每一对，对于一个未知样本每一个分类器都有一个选票，其结果是具有选票最多的类别。

表2 one-against-all的实验结果

| 组块类型 | Precision | Recall | FB1 |
|---|---|---|---|
| BAP | 76.60 | 83.72 | 80.00 |
| BDP | 100.00 | 100.00 | 100.00 |
| BMP | 94.74 | 97.30 | 96.00 |
| BNP | 87.30 | 88.00 | 87.65 |
| BNS | 83.33 | 83.33 | 83.33 |
| BTP | 77.78 | 77.78 | 77.78 |
| BVP | 87.33 | 89.73 | 88.51 |

表2给出了使用one-against-all多分类器扩展方法对7种汉语组块类型识别的实验结果。

### 4.3 利用分类器的函数值

在实验1中，会出现两个或多个分类器对于一个未知类别的样本给出+1的标识，即对两个或多个类别都投了选票，同样也有可能出现没有一个分类器对于一个未知类别的样本给出+1的标识的情况，即没有对一个类别投选票情况。

上述情况都难于决策当前位置的组块类型边界标记应该属于什么类别。考虑到SVMs是通过在训练过程中寻找一个间隔最大的超平面，而间隔定义为这个平面和与这个平面最近的训练样本(支持向量)之间的距离，所以可以基于这样的直觉：这个距离越大则这个分类结果越可信。进而进行下面的

实验。

实验2：设分类器 $c_i$ 对某未知类别样本的输出值为 $g_i(x)$，则该样本 $x$ 到分类器 $c_i$ 所构建的最优超平面的距离为 $d(w_i,b_i;+x)=\frac{|g_i(x)|}{\|w_i\|}$，通过比较多个分类器针对该未知类别样本的函数值 $g_i(x)(1\leqslant i\leqslant 15)$，从中选取最大的 $g_i(x)(1\leqslant i\leqslant 15)$所对应的类别。即，

$$c^*=\arg\max_{c_i}|g_i(x)| \tag{3}$$

表3 投票选举时利用分类器的函数值

| 组块类型 | Precision | Recall | FB1 |
|---|---|---|---|
| BAP | 80.00 | 83.33 | 81.3 |
| BDP | 100.00 | 100.00 | 100.00 |
| BMP | 94.74 | 97.30 | 96.00 |
| BNP | 83.78 | 86.11 | 84.93 |
| BNS | 85.54 | 81.99 | 83.73 |
| BTP | 77.78 | 77.78 | 77.78 |
| BVP | 87.33 | 89.73 | 88.51 |

利用分类器的函数值的实验2结果表明，BAP 组块识别查准率提高了3.4%，查全率却下降了0.39%，BNS 组块识别查准率提高了2.21%，查全率下降了1.34%，而 BNP 组块识别查准率和查全率都略有下降，其他类型的组块识别没有变化，可见投票选举过程中利用分类器函数值并不能使所有类型的组块识别有稳定的提高。

### 4.4 Sigmoid 函数

投票选举过程中利用分类器的函数值并不能使所有类型的组块识别有稳定的提高，其原因是分类器的函数值并不能保证整个分类过程的一致性，因此所作出的决策也不具有公平性，一个分类器在做全局决策的一个局部决策时，分类结果必须考虑到全面分类。因此我们希望一个分类器的输出应该是一个校准后验概率才能够进行后处理，即分类器应该产生一个非常有实际应用意义的后验概率 $p(class/input)$。

Sigmoid 方法是 John C. Platt 在1999提出的[14]，该方法使用一个参数模型来直接拟合后验概率 $p(y=1|g)$，而不是估计类别条件密度，通过调整模型的参数来给出最好的后验概率输出。

Sigmoid 使用一个参数模型来拟合最好的后验概率，这个参数模型的形式为：

$$p(y=1|g)=\frac{1}{1+\exp(Ag+B)} \tag{4}$$

参数 $A$ 和 $B$ 在训练了的 SVMs 基础上得到，即从训练集 $(g_i,t_i)$中使用极大似然估计得到。这里 $g_i=g(x_i)$，目标概率 $t_i$ 定义为：

$$t_i=\frac{g_i+1}{2}$$

通过最小化训练数据的负对数似然估计，即交叉熵错误函数来计算参数 $A$ 和 $B$：

$$\min-\sum_i t_i\log(p_i)+(1-t_i)\log(1-p_i)$$

其中 $p_i=\frac{1}{1+\exp(Ag_i+B)}$

在最小优化过程中需要解决两个问题：Sigmoid 训练集的选择，避免过度适应这个训练集的方法。

因为交叉检验较其他方法为 Sigmoid 提供了更大的训练集，我们通过交叉检验来估计参数 $A$ 和 $B$，所以估计出的参数 $A$ 和 $B$ 具有较小的偏差。将语料平均地分成10份，用10份语料中的9份组合训练了10个 SVMs，剩下的1份语料作为测试集，10个向量机在测试语料中的输出 $g_i$ 作为 Sigmoid 的训练集。由于本文定义了15种汉语组块类型边界标记，使用了多分类器 SVMs，因此在此基础上训练了15对 Sigmoid 函数的参数，表4给出了每一对参数值。

表4 交叉检验估计出的 Sigmoid 参数值

| 分类器 | A | B |
|---|---|---|
| 0 | −1.22345 | 3.54583 |
| 1 | −1.6878 | 2.52206 |
| 2 | −6.58168 | 0.204208 |
| 3 | −4.70771 | 1.77015 |
| 4 | −1.42426 | 1.58767 |
| 5 | −3.58818 | −0.354963 |
| 6 | −3.05942 | 0.0291156 |
| 7 | −3.97426 | 1.6672 |
| 8 | 0 | 9.97511 |
| 9 | −2.80512 | 4.7527 |
| 10 | −5.01868 | 0.421334 |
| 11 | −4.86146 | 0.481041 |
| 12 | −1.50136 | 0.762406 |
| 13 | −1.94256 | −0.306017 |
| 14 | −1.20067 | −0.204161 |

即使交叉检验为 Sigmoid 提供了没有偏差的训练数据，Sigmoid 也会过度适应，由于一些类别中存在能够从所有的反例中线性分开的极少正例，为了使 Sigmoid 函数最大可能地拟合这些 SVMs，即使给反例再次赋值，也会简单地给 $A$ 赋一个很大的负数，从而当检验集合可以正确分开时，对于无限陡 Sigmoids 会具有无限多的解决办法。因此当观察到一个正例 $g_i$ 时，并不使目标概率 $t_i=1$，而是假设在样本数据外有可能具有相反的类别的 $g_i$。因此对某一 $\varepsilon_+$，给正例 $g_i$ 赋予的目标概率为 $t_i=1-\varepsilon_+$，同样赋予一个负例的目标概率为 $t_i=\varepsilon_-$。

通过表4得到的15对 Sigmoid 函数参数值，我们将15个 SVMs 的输出映射成一个校准后验概率，在投票选举时选取最大概率值对应的类别，即：

$$c^*=\arg\max_{c_i}p_i=\arg\max_{c_i}\{1/(1+\exp(A_ig_i+B_i))\} \tag{5}$$

表5 投票选举时利用 Sigmoid 输出的后验概率

| 组块类型 | Precision | Recall | FB1 |
|---|---|---|---|
| BAP | 83.33 | 83.33 | 83.33 |
| BDP | 100.00 | 100.00 | 100.00 |
| BMP | 94.74 | 97.30 | 96.00 |
| BNP | 88.19 | 89.60 | 88.89 |
| BNS | 86.11 | 89.86 | 87.94 |
| BTP | 77.78 | 77.78 | 77.78 |
| BVP | 88.67 | 91.10 | 89.86 |

表5是投票选举时利用 Sigmoid 输出的后验概率来进行决策的实验，结果表明 BAP 组块的准确率提高了3.33%，BNP 组块的准确率提高了4.41%，BNS 组块的准确率提高了2.78%，BVP 组块的准确率提高了1.34%个百分点。

**结束语** 本文提出了一个 SVM＋Sigmoid 的汉语组块识别方法，并对模型中的多分类支持向量机，输入特征，Sigmoid 函数参数训练进行了详细的讨论。提出利用 Sigmoid 函数的后验概率，对多个类别投票数相等或没有对一个类别进行投票时获得合理的决策。实验结果也证明了 SVM＋Sigmoid 的方法比单一的支持向量机性能更高。

本文定义了7种汉语组块类型：基本副词短语(BDP)，基本形容词短语(BAP)，基本数量短语(BMP)，基本时间短语

(BTP),基本处所短语(BNS),基本名词短语(BNP)和基本动词短语(BVP)。一种组块类型有两种边界标记,加上不属于任何组块类型的边界标记O,一共定义了15种汉语组块类型边界标记。

本文仅使用了词、词性信息。在将来的研究过程中,我们将结合语义、搭配、共现的知识,以取得更好的汉语组块识别效果。

## 参 考 文 献


1 张昱琪,周强.汉语基本短语的自动识别.中文信息学报,2002,16(6):1~8.
2 朱德熙.语法讲义.商务印书馆,1982
3 Kudo T, Matsumoto Y. Chunking with Support Vector Machines. ACL,2001
4 周强,孙茂松,黄昌宁.汉语句子的组块分析体系.计算机学报,1999,22(11):1158~1165
5 赵军,黄昌宁.基于转换的汉语基本名词短语识别模型.中文信息学报,1998,13(2):1~7
6 Kudo T, Matsumoto Y. Use of Support Vector Learning for Chunk Identification, CoNLL,2000
7 Nakagawa T, Kudoh T, Matsumoto Y. Unknown word guessing and part-of-speech tagging using support vector machines. In: Proc. of the Sixth Natural Language Processing Pacific Rim Symposium,2001. 325~331
8 Yamada H, Kudoh T, Matsumoto Y. Japanese named entity extraction using support vector machines (in Japanese), In IPSJSIG Notes NL-142~17,2001
9 Joachims T. Learning to Classify Text using Support Vector Machines. Dissertation, Kluwer,2002
10 Abney S. Parsing by chunks. In Principle-Based Parsing Kluwer Academic Publishers,1991
11 Erik F, Sang T K, Buchholz S. Introduction to the CONLL-2000 Shared Task: Chunking[a], 2000. In: Proc. of CONLL-2000and LLL-2000. 127~132
12 Vapnik V N, Chervonenkis A Y. Theory of Pattern Recognition. (in Russian)Nauka, Moscow,1974
13 Introduction to Support Vector Machines. Dustin Bosw-ell. Aug. 2002
14 Platt J C. Probabilities for SV machines. In: A. J. Smola, P. L. Bartlett, B. Scholkopf, D. Schuurmans, eds. Advances in Large Margin Classifiers, MIT Press,2000. 61~71


(上接第133页)

表1 医疗分析信息表

| Patients | Condition1 | Condition2 | Decision |
|---|---|---|---|
| $x_1$ | $(0.2,0.3,0.2)/L+(0.8,0,8,0.2)/S$ | $(0.4,0.3,0.2)/H+(0.9,0.9,0.7)/P$ | $(0.3,0.2,0.2)/A+(0.9,0.7,0.6)/B$ |
| $x_2$ | $(0.3,0.3,0.2)/L+(0.9,0,7,0.7)/S$ | $(0.4,0.3,0.1)/H+(0.8,0.8,0.2)/P$ | $(0.8,0.8,0.2)/A+(0.9,0.7,0.6)/B$ |
| $x_3$ | $(0.8,0.8,0.3)/L+(0.3,0.2,0.2)/S$ | $(0.9,0.7,0.5)/H+(0,9,0.7,0.6)/P$ | $(0.9,0.7,0.4)/A+(0.8,0.8,0.3)/B$ |
| $x_4$ | $(0.9,0.6,0.7)/L+(0.9,0.8,0.5)/S$ | $(0.1,0.3,0.2)/H+(0.9,0.6,0.8)/P$ | $(0.9,0.7,0.6)/A+(0.3,0.2,0.2)/B$ |
| $x_5$ | $(0.2,0.3,0.3)/L+(0.9,0.6,0.8)/S$ | $(0.1,0.3,0.2)/H+(0.9,0.7,0.7)/P$ | $(0.1,0.2,0.2)/A+(0.4,0.1,0.2)/B$ |

这是一个F-模糊集属性信息系统表,在这个表中Condition 1和Condition2为条件属性,Decision为决策属性。

本文以考察有关特征和病症的关系来说明关于F-模糊集属性信息系统的知识获取方法。

由表1可以得到:

$$L=(0.2,0.3,0.2)/x_1+(0.3,0.3,0.2)/x_2+(0.9,0.6,0.7)/x_3+(0,9,0.6,0.7)/x_4+(0.2,0.3,0.3)/x_5$$

$$S=(0.8,0.8,0.2)/x_1+(0.9,0.7,0.7)/x_2+(0.3,0.2,0.2)/x_3+(0.9,0.8,0.5)/x_4+(0.9,0.6,0.8)/x_5$$

$$H=(0.4,0.3,0.2)/x_1+(0.4,0.3,0.1)/x_2+(0.9,0.7,0.5)/x_3+(0.1,0.3,0.2)/x_4+(0.1,0.3,0.2)/x_5$$

$$P=(0.9,0.9,0.7)/x_1+(0.8,0.8,0.2)/x_2+(0.9,0.7,0.6)/x_3+(0.9,0.6,0.8)/x_4+(0.9,0.7,0.7)/x_5$$

$$A=(0.3,0.2,0.2)/x_1+(0,8,0.8,0.2)/x_2+(0.9,0.7,0.4)/x_3+(0.9,0.7,0.6)/x_4+(0.1,0.2,0.2)/x_5$$

$$B=(0.9,0.7,0.6)/x_1+(0.9,0.7,0.6)/x_2+(0.8,0.8,0.3)/x_3+(0.3,0.2,0.2)/x_4+(0.4,0.1,0.2)/x_5$$

这是一种病例对诸特征和病症的支持度的F-模糊集合表示。若要考察某些特征对病症的必然或可能的关联程度用式(3)和(4)即可得到。下面仅就特征$L$和$H$对于病症$A$必然和可能的关联程度进行计算。

$$L\cap H=(0.2,0.3,0.2)/x_1+(0.3,0.3,0.1)/x_2+(0.9,0.6,0.5)/x_3+(0.1,0.3,0.2)/x_4+(0.1,0.3,0.2)/x_5$$

表示同时具有特征$L$和$H$的F-模糊集。利用式(3)和(4)可以得到:

$$FI((L\cap H)\subset A)=(0.8,0.7,0.5)$$

$$FJ((L\cap H)\#A)=(0.9,0.8,0.6)$$

**结论** 本文提出了F-模糊集的概念,给出了利用F-模糊集在信息系统中进行知识获取的方法。模糊集合是对经典集合的补充,解决了由经典集合难以表示和解决的问题。F-模糊集是对模糊集的补充,以解决在实际应用中直接利用模糊集难以解决的问题。文中对F-模糊集导出的矩阵表示对于理论研究及应用具有一定意义。利用矩阵表示和运算对F-模糊集属性信息系统规则约简和属性约简是本项目进一步研究的一个方面。

## 参 考 文 献


1 Zadeh L A. Fuzzy sets. Inform. and Control,1965, 8: 338~353
2 de Korvin A, McKeegan C. Knowledge acquisition using rough sets when membership values are fuzzy sets. J. Intelligent and Fuzzy Systems,1998, 6:237~244
3 Hong T P, et al. Learning a coverage set of maximally general fuzzy rules by rough sets. Experts Systems with Applications, 2000, 19: 97~103
4 曾黄麟.粗集理论及其应用.重庆:重庆大学出版社,1996
5 张文修.粗糙集理论与方法.北京:科学出版社,2001
6 闫德勤,迟忠先.一种实值属性信息系统的粗集约简方法.小型微型计算机系统,2003,24(3):517~519
7 张文修.模糊数学引论.西安:西安交通大学出版社,1991