文章编号:1008-0570(2007)10-3-0205-03

基于SVM的词频统计中文分词研究

Study on Chinese word segmentation based on statistic and SVM

(中南大学)朱小娟 陈特放
ZHU XIAOJUAN CHEN TEFANG

摘要:本文详细介绍SVM(支持向量机)在词频统计中文分词中的应用。可将输入的连续字串进行分词处理,输出分割后的汉语词串,一般为二字词串,并得到一个词典。词典中不重复地存储了每次处理中得到的词语,以及这些词语出现的频率。选用了互信息原理进行统计。并采用SVM算法,分词的准确性与传统相比有了很大的提高,并具有一定的稳定性。

关键词:中文分词;词频统计;互信息;支持向量机
中图分类号: TP393 **文献标识码:**B

Abstract:The paper introduces the application of SVM in Chinese word segmentation, which is based on statistic the frequency of the word. Through the system, continuous character bunch input can be segmented, and then the cut apart word bunch output can be gotten, the cut apart word bunch usually is two character word bunch, and one dictionary can be gotten. The dictionary stores word and the frequency that the word appears in these disposal tests. The segmentation system selects Mutual Information to statistic. Use SVM, the veracity of segmentation was better than the traditional method, and is of high stability.

Key words:Chinese word segmentation,Statistic the frequency of the word,Mutual Information,SVM

技术创新

引言

统计学习理论是一种专门研究有限样本条件下机器学习规律的理论。该理论针对小样本统计问题建立了一套新的理论体系。统计学习理论的一个核心概念就是VC维概念,它是描述函数集或学习机器的复杂性或者说是学习能力的一个重要指标。在这一理论基础上发展了一种新的通用学习方法——支持向量机(SVM),已初步表现出许多优于已有方法的性能。支持向量机方法是建立在统计学习理论的VC维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力。目前,SVM算法在模式识别、回归估计、概率密度函数估计等方面都有应用。并表现出其较高的准确性。

词是最小的能够独立活动的有意义的语言成分,是自然语言处理系统中重要的知识载体与基本操作单元。中文分词就是由计算机自动识别文本中的词边界的过程,它是中文信息处理最重要的预处理。然而到目前为止,还没有真正成熟实用的中文分词系统面世,这成为严重制约中文信息处理发展的瓶颈之一。

在此背景下, 笔者研究了基于SVM的词频统计中文分词技术,介绍了词频统计分词技术的统计原理,并采用SVM进行分词处理。

1 SVM学习算法

SVM是根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力(Generalization Ability)。它通过事先选择的非线性映射Φ将输入向量映射到一个高维特征空间Z,在这个空间构造最优分类超平面。这样可使原始空间非线性可分的问题变为高维空间中线性可分的问题。

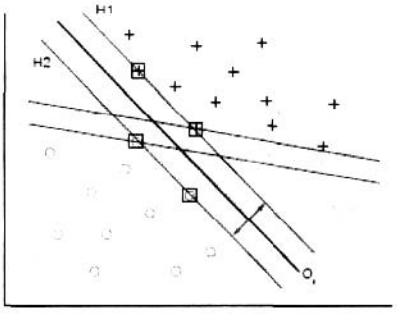

图1 数据点集的超平面

SVM的基本思想可用图1的两维情况来说明。图中,十字和空心点代表两类样本,Oi为分类线,H1、H2分别为过各类中离分类线最近的样本且平行于分类线的直线,它们之间的距离叫做分类间隔。所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为0),而且使分类间隔最大。分类线方程为$x \cdot w + b = 0$,对它进行归一化,使得对线性可分的样本集$(x_i, y_i), i = 1,\ldots, n$, $x \in R^d$, $y \in \{+1,-1\}$,满足公式1

$$y_i\left[(w.x_i)+b\right]-1 \geq 0, i = 1,\ldots,n \qquad \text{公式(1)}$$

此时分类间隔等于$2/\|w\|$,使间隔最大等价于使$\|w\|2$最小。满足条件(4)且使$\frac{1}{2}\|w\|^2$最小的分类面就叫做最优分类面,H_1、H_2上的训练样本点就称作支持向量。利用Lagrange优化方法

朱小娟:硕士
基金项目:国家自然科学基金资助项目(60674003)

可以把上述最优分类面问题转化为其对偶问题,而在最优分类面中采用适当的内积函数 K(xi,xj)就可以实现某一非线性变换后的线性分类,相应的分类函数为

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n} a_i^{-} y_i K(x_i,x)+b^{-}\right) \qquad (2)$$

式中求和实际只对支持向量进行。b* 是分类阈值,若 f (x) >0,则 x 属于类别 α,否则就不属于该类。

但标准 SVM 的输出值并不是一个全局的概率值,难以引入如 Bayes 决策等机制对多值分类问题进行优化处理。Platt 利用 sigmoid 函数实现了 SVM 算法的后处理,可以将 SVM 的输出值映射为概率值。Sigmoid 函数的转换形式如下:

其中,f 为标准 SVM 的输出结果,P(y =1l f) 表示在输出值 f 的条件下分类正确的概率。A 和 B 是函数中的参数值,其优化策略可通过解如下一个最大似然问题来解决:

$$P(y=1/f)=\frac{1}{1+\exp(Af+B)} \qquad (3)$$

其中,f 为标准 SVM 的输出结果,P(y =1l f) 表示在输出值 f 的条件下分类正确的概率。A 和 B 是函数中的参数值,其优化策略可通过解如下一个最大似然问题来解决:

$$F(z)=\min_{z=(A,B)}\left(-\sum_{i=1}^{l}(t_i\log(p_i)+(1-t_i)\log(1-p_i))\right) \qquad (4)$$

其中,$p_i=\frac{1}{1+\exp(Af_i+B)}$,$f_i=f(x_i)$,

$$t_i=\begin{cases}\dfrac{N_+ +1}{N_+ +2} & \text{if } y_i=1\\ \dfrac{1}{N_- +2} & \text{if } y_i=-1\end{cases}$$

(i=1,2,…l,N+ 、N-为模型中的正例,反例数目)

通过这种转换,可以在全局的范围内对单个 SVM 分类器的输出进行校正,提高 SVM 分类器的适应性与准确性。

2 词频统计原理

从形式上看,词是稳定的字的组合,因此在上下文中,相邻的字同时出现的次数越多,就越有可能构成一个词。因此字与字相邻共现的频率能够较好地反映成词的可信度。这就是词频统计的基本原理,最常用的互信息原理。

2.1 互信息原理

在人们用语言进行交际活动时,语言成分的使用显示出一定的规律性,因此可使用统计方法对其进行研究统计,语言学采用概率论、数理统计以及信息论等数学工具来研究语言成分出现的概率和频率,从而揭示语言的统计规律。

定义 1:对有序汉字串 AB 汉字 AB 之间的互信息定义如公式(1)所示。

$$I(A,B)=\log_2\frac{P(A,B)}{P(A)P(B)} \qquad \text{公式(1)}$$

互现信息体现了汉字之间结合关系的紧密程度,当紧密程度高于某一个阈值时,便可认为此字组可能构成了一个词。其中,P(A,B)为汉字串 AB 联合出现的概率,P(A)为出现汉字串 A 的概率,P(B)为汉字串 B 出现的概率,它们在汉字字符串中出现的次数分别计为 n(A)、n(B)、n(AB),n 是词频总数,则有公式(2)。

$$P(A,B)=\frac{n(AB)}{n},\quad P(A)=\frac{n(A)}{n},\quad P(B)=\frac{n(B)}{n} \qquad \text{公式(2)}$$

互信息反映了汉字串 AB 间相关的程度。

(1)如果 $I(AB)\geq 0$,即 $P(AB)\geq P(A)P(B)$,则 AB 间是正相关的,随着增加相关度增加,如果大于给定的一个阈值,这时可以认为 AB 是一个词;

(2)如果 $I(AB)\approx 0$,即 $P(AB)\approx P(A)P(B)$,则 AB 间是不相关的;

(3)如果 $I(AB)<0$,即 $P(AB)<P(A)P(B)$,则 AB 间是互斥的,这时 AB 间基本不会结合成词。

3 基于 SVM 的词频统计中文分词系统

3.1 基于 SVM 的词频统计中文分词系统的功能

(1)通过词频统计,对文本(生语料)进行分词,尽可能达到较高的准确率和较高的运行效率。

(2)能够在初步分词后,采用 SVM 对歧义字段进行分割,从而达到更高的分词准确率。

(3)能够将分割出的词语存入一个词典中,并且同一词语不重复存储。

(4)能对词典中词语出现的频率进行累加,也就是说,能对同一个词语在不同的文档中出现的频率进行累加。

3.2 基于 SVM 的词频统计中文分词系统的设计

这个分词系统主要由四部分组成:预处理模块,词频统计模块,SVM 处理模块,词典生成模块。首先预处理阶段,利用显式和隐式的切分标记将待分析的文本切分成短的汉字串。然后通过词频统计模块,统计出单字出现的频率,已及相邻两字出现的频率,并计算出相应的统计信息。最后根据词频统计模块中得到的统计信息,用 SVM 处理模块对文档进行分割(对歧义字段进行分割),并将切分出的词语存入词典中,而且将词语出现的频率存入词典中。

4 基于 SVM 的词频统计中文分词系统的实现

4.1 预处理模块

预处理模块的实现是比较简单的,它的关键问题是汉字的编码和数据结构。在汉字编码中,每个汉字由两个字节组成,并且将最高位置"1"。这样要判断一个字符是否为汉字,就只要进行位运算就可以判断,从而做初步的处理。

4.2 词频统计模块

词频统计模块的实现算法是由单字出现频率统计和相邻 2 字出现频率统计两个方面组成。要计算相邻 2 字共现的频率,实现的方法就是每次从数组中取出两个相邻的汉字,让它们和后面的字符进行匹配,来计算它们共现的次数,而且对每次统计过的汉字进行标记。

4.3 SVM 处理模块

这个模块是系统实现的关键,在词频统计模块中只计算出单字出现和相邻 2 字共现的频率,根据这些信息可以初步判断出相邻 2 个字是否能组成两字词语。然后将其收录在字典中。

然而,在其中存在很多的歧义字段。它严重影响了分词的精度。在这里,我们就采用 SVM 来解决歧义切分的问题。例如

技术创新

在字段 AJB 中，AJ 可以组成词语，并且 JB 也可以组成词语。通过传统计算统计信息，无法对其进行分割，因此，采用 SVM 来解决。首先将每个歧义字段转化为 SVM 能处理的输入输出。例如：对于歧义字段：JS：$a_1\cdots a_i b_1\cdots b_m c_1\cdots c_n (i>0,m>0,n>0)$存在两种切分方案：

SEG1：$\underset{w_{11}}{a_1\cdots a_i}\ \underset{w_{12}}{b_1\cdots b_m}\uparrow_{pt1}\ \underset{w_{13}}{c_1\cdots c_n}$　　(正向切分方案)

SEG2：$\underset{w_{11}}{a_1\cdots a_i}\uparrow_{pt2}\ \underset{w_{21}}{b_1\cdots b_m}\ \underset{w_{22}}{c_1\cdots c_n}$　　(反向切分方案)

其中，$w_{11}, w_{12}, w_{13}, w_{21}, w_{22}$ 均为词，pt1,pt2 分别对应 $b_m c_1$ 与 $a_i b_1$ 之间的位置。我今年将每个歧义字段表示为一个二维向量，记为<Ipt1,Ipt2>。Ipt1 表示正向切分断点处两字 $b_m c_1$ 的互信息值，Ipt2 表示正向切分断点处两字 $a_i b_1$ 的互信息值。互信息在前面已经介绍过。

4.4 词典生成模块

通过上面的词频统计模块和 SVM 处理模块。可将分割出的词语收录到词典中，并记录此词语出现的频率，以备下次判断时做参考。同时，在每次出现同样词语时，在字典中将其频率累加。

4.5 试验分析

从采用的 33.5M 的综合语料通过此分词系统处理。并选取 5643 个高频歧义字段训练后，得到的支持向量个数 1532 个，SVM 的核函数选择高斯核函数，参数选为 g=0.5，测试为开放测试。试验表明，在采用 SVM 时测试结果如下表所示：

表 1　用 SVM 算法的试验结果

类别	测试数据量	歧义字段个数	正确切分个数	正确率
经济	651K	1435	1310	91.2891%
文化	534K	1253	1143	91.2210%
教育	794K	1582	1430	90.3919%
平均				90.9367%

从试验结果可以看出，采用 SVM 来求解歧义切分具有较高的切分准确率，且对不同类的测试数据切分效果比较稳定。在传统词频统计分词系统中，没有采用 SVM，只是通过互信息统计来进行分割，其准确率比使用 SVM 算法的要低许多。

由于使用前面介绍的歧义字段表示方法，所得到的向量矩阵只有两维，由此构造的训练集和测试集规模小，所花费的训练和测试的时间也比采用高维向量所需的时间少。另一方面可看到选取较小数量的训练集，所花费的分类时间少且得到的测试结果比较稳定。

5　结论

本文介绍的是基于 SVM 的词频统计中文分词技术，采用互信息作为歧义字段的表示方法和分词的原理，并用 SVM 算法对歧义字段进行切分。试验证明，这种方法非常可行，有较高的准确率，并且有效提高了效率。同时我们发现对歧义字段的表示方法能较大的影响切分的准确率，因此可以考虑用其他方法，如 t 测试信息，N 元统计模型等。而且试验中发现被 SVM 错分的样本大多集中在最优超平面附近，因此下一步还可以研究是否可以采用和其他方法结合来提高切分准确率。

本文作者创新点：在传统的词频统计中文分词技术中，对于歧义字段采用 SVM 进行切分，极大的提高了切分准确性。通过对系统各模块的设计，对 SVM 核函数的选择和转换，并采用互信息对歧义字段进行表示，又提高了 SVM 和系统整体的效率。同时发现对歧义字段的表示方法能较大的影响切分的准确率，因此采用其他统计原理进行表示，以及提高 SVM 的准确度是下一步研究的方向。

参考文献

[1]Paul N.Bennett,Susan T.Dumais,Eric Horvitz. Probabilistic Combination of Text Classifiers Using Reliability Indicators:Models and Results. SIGIR '02,2002,207~214.

[2]朱辉，杨扬，颉斌，封筠. SVM 在小字符集手写体汉字识别中的应用研究[J]微计算机信息，2004，4：74~75.

[3]李蓉，刘少辉，叶世伟，史忠植.基于 SVM 和 KNN 结合的汉语交集型歧义切分方法[J].中文信息学报，2002，15(6)：13~18.

[4]李衍，朱靖波，姚天顺.基于 SVM 的中文组块分析[J].中文信息学报，2003，18(2):1~6.

作者简介：陈特放(1957-)，男(汉族)，湖南涟源人，博士生导师，博士，主要研究方向：电力机车与故障诊断，交通信息工程及控制，计算机应用；朱小娟(1982-)，女(汉族)，湖南怀化人，硕士，主要研究方向：计算机信息处理与控制。

Biography:Chen tefang(1957-), male (the Han nationality), Hunan Province, Central South University, Ph.D,Professor, Research area: Electric train and fault diagnosis,traffic information and control, application of compute; Zhu xiaojuan(1982-), female(the Han nationality),Hunan Province, Central South University, M.S, Research area: The disposal and control of compute information.

(410075 长沙　中南大学信息科学与工程学院)朱小娟　陈特放

通讯地址：(410075 长沙　湖南长沙中南大学铁道校区十四舍 103)朱小娟

(收稿：2007.7.23)(修稿日期：2007.9.25)

技术创新