直推式支持向量机在 Web 信息抽取中的应用研究

肖建鹏,张来顺,任 星
XIAO Jian-peng,ZHANG Lai-shun,REN Xing

中国人民解放军信息工程大学 电子技术学院,郑州 450004
Institute of Electronic Technology,the PLA Information Engineering University,Zhengzhou 450004,China
E-mail:betret@sohu.com

XIAO Jian-peng,ZHANG Lai-shun,REN Xing.Web information extraction based on Transductive Support Vector Machine.Computer Engineering and Applications,2009,45(2):147-149.

Abstract: Transductive Support Vector Machines(TSVM) classify the new data vector based on the information only related to this data vector.This paper proposes a Web information extraction method based on TSVM and extract Web information with the classify angle.It needs far less tagged samples to carry out classify mark a lot of untagged samples and complete Web information extraction by classified way.The results show that TSVM can be used in Web information extraction.

Key words: Web information extraction;classification learning;Transductive Support Vector Machine(TSVM)

摘 要:直推式支持向量机是一种直接从已知样本出发对特定的未知样本进行识别的分类技术。在分析直推式支持向量机分类原理的基础上,提出一种基于直推式支持向量机的 Web 信息抽取方法,直接从分类的角度抽取 Web 信息。只需要提供少量标记样本就可以实现对大量未标注样本的分类标注,从而以分类的方式完成 Web 数据抽取任务。实验结果表明,使用这种方法进行 Web 信息抽取是有效性。

关键词:Web 信息抽取;分类学习;直推式支持向量机
DOI:10.3778/j.issn.1002-8331.2009.02.043 **文章编号**:1002-8331(2009)02-0147-03 **文献标识码**:A **中图分类号**:TP311

1 引言

随着 Internet 及其技术的迅速发展,Web 已经成为当今最庞大的信息库。然而 Web 页面中通常含有很多用户并不关心的信息,如广告链接、导航栏和版权信息等,如何从 Web 页面中抽取出有用的信息已经成为当前信息领域的研究热点之一。

由于传统的 Web 信息抽取技术上的不足, 人们开始在 Web 信息抽取系统中引入数据挖掘技术和机器学习方法,使抽取系统具备自适应和自动学习的能力。支持向量机(Support Vector Machines,SVM) 技术作为统计学习理论的重要发展成果,开始被应用到 Web 信息抽取中。本文主要研究直推式支持向量机(Transductive Support Vector Machine,TSVM)在 Web 信息抽取中的应用。

2 支持向量机理论

SVM 是在结构风险最小化原理(Structural Risk Minimization,SRM)的基础上发展起来的机器学习方法。按照不同的推导理论,通常分为归纳式支持向量机(ISVM)和直推式支持向量机(TSVM)[1]。SVM 最初被用于解决模式识别问题,其基本思想是: 通过用非线性映射 φ 将输入空间变换到一个高维空间,在这个高维空间中寻找输入变量和输出变量之间的一种非线性关系[2]。注意到算法仅使用高维空间中的内积,通过引入核函数,高维空间的内积运算就可用原空间中的函数来实现,甚至没有必要知道 φ 的形式。通过采用适当的核函数就可实现非线性变换后的线性分类,而计算的复杂度则没有增加,从而在一定程度上避免了维数灾难问题。

2.1 归纳式支持向量机

传统的 SVM 是从训练样本出发,在全局空间对问题求解。因此,归纳式 SVM 是利用训练样本集来建立分类决策函数,最终得到最优分类超平面(决策函数)可以表示为:

$$f(x)=sign\{\sum_{a_i>0} a_i y_i K(x_i,x)-b_0\}$$

其中 K 为核函数。通常 $a_i>0$ 对应的样本点被称为支持向量(Support Vector,SV)。

2.2 直推式支持向量机

在传统的归纳式向量机中, 训练 SVM 学习机需要大量经过标记的样本,而正确标记的样本是很难大量获取的。如果能够把未标记的样本的特征加入到已标注样本中去,就可以弥补归纳式 SVM 带来的不足。直推式 SVM 正是基于这种思想的 SVM 算法[3-4]。

TSVM 是一种不依赖于推广性思想的经验推理。由于其从

作者简介:肖建鹏(1979-),男,硕士研究生,研究方向为 Web 挖掘、信息抽取;张来顺(1963-),男,教授,硕士生导师,研究方向为计算机应用技术;任星(1982-),硕士研究生,研究方向为数据库安全。

收稿日期:2007-12-27 **修回日期**:2008-03-17

特殊到特殊的推理,难以直接进行客观验证。因此,直到现在才开始得到人们研究的重视,但它已经在一些领域中(例如生物基因选择,数字识别)[5]取得初步成果,甚至表现出比传统的归纳式 SVM 具有更好的性能。直推式 SVM 的分类决策函数是建立在训练集 S_{train} 和测试集 S_{test} 的基础上[6],即

$$F_L=L(S_{train},S_{test})$$

$$W_{HERE}:S_{train}=(x_1^*,y_1^*),(x_2^*,y_2^*),\cdots,(x_n^*,y_n^*)$$

根据支持向量机理论,分类超平面满足最优分类超平面的条件为

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

$$\min_w \phi(w)=\|w\|^2$$

为了保证输入向量在线性不可分的情况下,允许错分样本的存在,引入了松弛变量 ξ_1。然后利用 Lagrange 优化方法,引入 Lagrange 乘子 α_i,$i=1,2,\cdots,l$ 后,问题式(1)就转变为

$$Minimiz\ \ over(y_1^*,\cdots,y_k^*,w,b,\xi_1,\cdots,\xi_n,\xi_1^*,\cdots,\xi_k^*)$$

$$\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i+C^*\sum_{j=1}^{k}\xi_j^* \tag{2}$$

$$\text{s.t.}\ \ \forall_{i=1}^{n}:y_i[w\cdot x_i+b]\geqslant 1-\xi_i$$

$$\forall_{j=1}^{k}:y_i[w\cdot x_i^*+b]\geqslant 1-\xi_j^*$$

$$\forall_{i=1}^{n}:\xi_i\geqslant 0$$

$$\forall_{j=1}^{k}:\xi_i\geqslant 0$$

可以通过控制惩罚因子 C 和 C^* 的大小,来调节误分类样本对分类超平面的影响,测试集样本就是通过包含惩罚因子 C^* 的项来影响分类超平面的。

3 基于 TSVM 的 Web 信息抽取

网页作为一种特殊的文档,从其结构来看可分为内容信息和格式信息。内容信息是指通过浏览器呈现给用户的信息,主要以文本为主,在内容信息中被标记分割开的块叫做信息片断。格式信息是指用于表示和解释内容信息的信息。

把网页中不同的信息片断归纳到不同的类别,将网页中有用数据和无用数据正确区分开,是以分类的角度来看待 Web 信息抽取问题的主要思想。传统的 Web 信息抽取需要用户大量的标记数据进行训练学习,从而导致自动化程度降低。TSVM 可以结合已标注的样本,将大量未标记的数据作用于 SVM 学习机中。因此,TSVM 是一种更适合于当今大数据量抽取的机器学习方法。

3.1 方法的总体框架

基于 TSVM 的 Web 信息抽取方法主要由网页采集、网页预处理、特征提取和数据抽取 4 个部分组成,整个方法的构架如图 1 所示。从图中可以看出,在 TSVM 的学习过程中,目标页面就开始对分类学习施加影响。

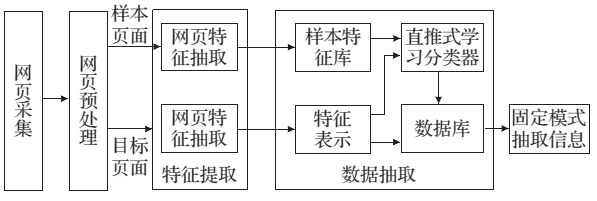

图 1 基于 TSVM 的 Web 信息抽取方法流程图

3.2 网页采集

网页采集模块主要负责从网络上搜集并下载原始的网页。获取的方法可以利用搜索引擎返回的网页地址作为输入,通过抓取网页程序 crawler 将待抽取网页抓取并存入本地网页数据库中。由于本文主要研究网页的数据抽取问题,因此这部分不作为本文重点描述。

3.3 网页预处理

网页预处理模块的主要目的是对所获得的样本网页和目标网页进行修复和解析处理。因为浏览器具有容错的能力,即使文档中的 HTML 标记不正确,Web 文档还是能正确展示出来,所以需要先将结构不完整或不规范的 HTML 文档转换成结构良好的 XHTML 文档并解析成 DOM 树结构。几乎所有的 HTML 文档都可以被解析成 DOM 树,树上的所有叶子节点都被认为是信息片断,每个信息片断对应着 HTML 文档中的一行文本,其中不仅包含内容信息而且还包含格式信息。

3.4 网页特征提取

成功的特征选择是将分类方法用于 Web 信息抽取的关键。因此,特征提取模块的主要作用是获取 DOM 树中信息片断的特征,从而将网页特征向量化。所获取的特征有以下几类:(1)位置特征:信息片断所在的位置具有相对的确定性,DOM 树中具有层次结构的路径可以作为信息片断的“坐标”。(2)上下文特征:是指网页中信息片断的前后信息(之前叫做前引导词,之后叫做后引导词)往往与信息片断有一定的关系,通常具有提示作用。(3)一般特征:是指不同信息片断之间的本质区别,比如时间用数字表示而姓名用字母表示等。(4)可视特征:HTML 标记不仅可以用来组织内容,还可以用来表示网页的外观,如字体的大小和颜色、段落的长短等,在这里把这种视觉上的特征叫做可视特征[7]。

通过以上 4 种特征的描述,网页中的信息片断就可以用特征向量来表示,具体获取特征的方法是通过 DOM 树所生成的信息片断来构造的,过程如下:一般特征和可视特征在不同的网页中各不相同,检测即可以获得。上下文特征的获得只需要定位引导词,不需要更进一步的处理。位置特征的获得相对来说比较复杂,可以通过以下的方法获得:

```
Check(C_Tag);// 检查当前标记
 if C_Tag==H_Tag // 当前标记是头标记
  Path=Add_Path(C_Tag);// 将头标记加入到路径标记中
   if Pre_H_Tag==H_Tag // 头标记的前一个标记是头标记
    Path=Add_Path(Pre_H_Tag);// 加入到路径标记中
   else Path_Value=+1;// 头标记的前一个标记是尾标记,修改路径序号值
  else Path=Remove_Last();// 当前标记是尾标记,移除末端标记
if Pre_T_Tag=T_Tag // 尾标记的前一个标记是尾标记
 Path=Remove_Last();// 移除末端标记
  Pre_T_Tag=C_Tag;// 重置前一个标记
```

3.5 数据抽取

数据抽取模块是整个方法的核心部分。为了在目标网页中抽取出用户所需的信息,可以通过使用 SVM 算法对一般特征、可视特征、上下问特征和位置特征这 4 类特征进行训练以达到将网页中的信息片断分类标注的目的。传统的 SVM 方法仅仅通过标注的样本来分类未标注的样本,然而对网页这种包含有大量特征样本的数据来说,通过已标注的样本来分类未标注样本其效率是十分低的。因此,本文使用 TSVM 作为分类方法的

核心部分。

从图 1 可以看出,在 TSVM 学习期间,样本页面就已经对其施加影响。因此,TSVM 学习的结果中就已包含样本页面的数据特征。经过学习后,TSVM 就会产生一个分类决策函数,即最优分类超平面。特征化的目标样本经过 TSVM 分类器对页面中的信息片断进行分类标注,最后将标注好的数据存入数据库中。

为了尽可能的分类出待抽取数据,以提高数据抽取效率,TSVM 试图寻找最优分类超平面。而 TSVM 的有效学习是非常重要的,其具体的学习过程如下:

步骤 1　指定参数 C 和 C^*,使用归纳式学习对有标识样本进行一次初始学习,得到一个初始分类器 F,并按照某个规则指定一个未标识样本中的正标签样本数 N;

步骤 2　用初始分类器 F 对未标识样本进行分类,根据对每一个未标识样本的判别函数输出,对输出值最大的 N 个未标识样本暂时赋正标识值,其余的赋负标识值。并指定一个临时影响因子 C^*_{temp};

步骤 3　对所有样本重新训练,对新得到的分类器 F_1,按一定的规则交换一对标签值不同的测试样本的标签符号,使得优化问题(2)中的目标函数值获得最大下降,这一步骤反复执行,直到找不出符合交换条件的样本对为止;

步骤 4　均匀地增加临时影响因子 C^*_{temp} 的值并返回到步骤 3,当 $C^*_{temp}>C^*$ 时,算法结束,并输出结果。

当 TSVM 终止学习后,就可以使用学习得到的分类器对目标网页中的信息片断进行正确的分类标注。TSVM 在学习的过程中,结合大量无标签样本信息的同时反复调整无标签样本对 TSVM 学习机的影响,以追求对无标签样本的最小分类误差。因此,基于 TSVM 的 Web 信息抽取方法在数据分类的精确度上会得到一定程度的增强,进而提高数据抽取的准确性。

经过 TSVM 分类器后,DOM 树中的信息片断被标注为不同的类别存储在数据库中。当用户查询所需信息时,根据所指定的条件定位到数据库中,取出数据返回给用户。

4　实验结果和分析

4.1　实验数据和环境

使用一台 PⅣ 2.0 GHz,512 MB 内存的 PC 机进行实验,系统的实验环境为 Windows XP,利用 VC6.0 实现类似微软浏览器 IE 的程序界面。系统通过网页爬虫程序从网站获取页面,根据样本页面学习获取信息片断的特征,对目标页面进行处理获取页面的信息片断及其特征,采用 TSVM 分类器对各个信息片断进行分类标注,将结果存储到 XML 文档和数据库中,信息抽取结果使用 Web 表示,返回 XML 和数据库方式的信息查询。

实验数据选取搜狐、新浪、网易、雅虎 4 个网站体育版的体育新闻作为实验对象进行评测,这些网页结构差别大,有助于验证方法的性能,实验的目的是验证 TSVM 在 Web 信息抽取中的可行性和数据抽取质量。

4.2　评估标准

信息抽取能性的主要评价指标是召回率(#Recall)和准确率(#Precision),召回率等于系统正确抽取的结果(#Real)占所有可能正确结果(#True)的比例;准确率等于系统正确抽取的结果(#Real)占所有抽取结果(#Total)的比例。为了综合评价数据抽取的性能,通常还计算召回率和准确率的加权几何平均值,即 F 指数,它的计算公式如下[8]:

$$F=(\beta^2+1)\times Precision\times Recall/\beta^2\times Precision+Recall$$

其中,β 是召回率和准确率的相对权重。β 等于 1 时,二者同样重要;β 大于 1 时,准确率更重要一些;β 小于 1 时,召回率更重要一些。

4.3　结果与分析

当取 β=1 时,选取 4 个网站中关于欧洲冠军杯的新闻进行测试,所得结果如表 1 所示。

表 1　4 个网站冠军杯信息抽取测试结果

WebSite	#Total	#Real	#True	#Recall	#Precision	#F
搜狐体育	195	182	182	0.933	1.00	0.965 4
新浪体育	283	231	231	0.816	1.00	0.898 7
网易体育	331	307	307	0.927	1.00	0.962 1
雅虎体育	359	336	336	0.936	1.00	0.966 9

由表 1 可知,本文所提出的抽取方法能够很好地获取分类决策函数并准确的抽取出所需信息。在对没有抽取到的网页做出进一步分析发现,抽取失败的主要原因是由于网页中新闻信息片断的特征有所不同,因此一些的新闻信息没有被正确特征向量化,从而引起信息片断被忽略,最终导致召回率的下降。但是,将信息抽取技术和机器学习中的分类方法相结合,一方面减少信息片断因内部的结构变化带来的抽取失败,另一方面避免过多的人工干预,实现数据抽取的自动化。

5　结束语

本文提出一种基于 TSVM 的 Web 信息抽取方法。以分类的角度抽取 Web 信息,只需要用户提供少量的信息就可以完成抽取任务。通过对 Web 上关于欧洲冠军杯新闻的测试表明:该方法在没有用户大量参与的情况下,用自动提取的特征作为分类依据,抽取信息仍然保持高度的准确性,对于结构变化较大网页具有更好的健壮性。

参考文献:

[1] 许建华,张学工.统计学理论基础[M].北京:电子工业出版社,2004.

[2] Vapnik V.The nature of statistical learning theory[M].New York:Springer-Verlag,2000.

[3] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习[J].软件学报,2003,14(3):451.

[4] Joachims T.Transductive inference for text classification using support vector machines[C]//Proceeding of the 16th International Conference on Machine Learning.San Francisco:Morgan Kaufmann,1999:200-209.

[5] Thorsten J.Transductive inference for text classification using support vector machines[C]//Proc of International Conference on Machine Learning.San Francisco,CA,USA:Morgan Kaufmann,1999:200-209.

[6] Nikola K,Shaoning P.Transductive support vector machines and applications in bioinformatics for promoter recognition,letters and reviews[J].Neural Information Processing,2004,3(2):31-38.

[7] Yu S P,Cai D,Wen J R,et al.Improving pseudo-relevance feedback in Web information retrieval using Web page segmentation[EB/OL].http://research.microsoft.com/research/pubs/view.aspx?type =Technica1}20Report&id=6322002.

[8] 李保利,陈玉忠,俞士汶.信息抽取研究综述[J].计算机工程与应用,2003,39(10):1-5.