

## 基于潜在链接分析的 FTSM 网页分类

王 晔, 黄上腾

(上海交通大学计算机系, 上海 200030)

**摘要:** 回顾了模糊直推式模糊支持向量机 (FTSM) 的不足, 并提出了一种基于潜在链接分析并结合网页权重信息的 FTSM 网页分类方法。新方法提高了分析网页超链接信息的效率, 避免了经验参数的影响, 充分考虑了网页权重的贡献。试验表明, 提出的方法在网页基准测试数据上取得了优于 FTSM 的分类效果。

**关键词:** 网页分类; 直推式支持向量机; 模糊; 潜在链接分析; 网页权重

## Latent Link Analysis-based FTSM for Hypertext Classification

WANG Ye, HUANG Shangteng

(Department of Computer, Shanghai Jiaotong University, Shanghai 200030)

**【Abstract】** This paper analyzes the deficiencies of the FTSM. To overcome these shortcomings, it proposes a latent link analysis-based FTSM (LLA\_FTSM) for hypertext classification. The new method takes the information of weights of hypertexts into consideration. Experimental results show that the LLA\_FTSM performs well.

**【Key words】** Hypertext classification; TSVM; Fuzzy; Latent link analysis; Weight of hypertext

网页是具有带有结构信息并说明链接关系的文本, 与纯文本相比, 网页的信息量更大、网页与网页之间的关系更密切, 但同时也比纯文本分类问题更加难处理。

本文是在刘宏、黄上腾提出的用于网页分类的直推式模糊支持向量机(FTSM)<sup>[1]</sup>工作的基础上进一步深入研究, 着重讨论了 FTSM 在超链接分析和时间复杂度方面的不足, 并将网页权重信息结合到 FTSM 训练过程中, 使支持向量机分界面的位置更为合理。

## 1 对 SVM 和 FTSM 的回顾

传统的分类方法, 如朴素贝叶斯分类、神经网络等都在网页分类方面取得了一定的成果。基于支持向量机的网页分类是近年发展起来一种新方法, 下面简要回顾其原理和应用。

## 1.1 支持向量机

支持向量机(SVM)<sup>[2]</sup>是由 Vapnik 在基于结构风险最小化准则提出的一种分类方法, 它用于解决高维、小样本、非线性的机器学习问题。支持向量机的训练阶段通过最小化目标函数寻找最优分界面, 即

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_i \xi_i \quad (1)$$

$$\text{st. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

其中,  $w$  是最优决策面的方向向量,  $b$  是阈值,  $x$  是样本向量,  $y$  是类别标记,  $\xi$  是松弛变量;  $C$  是常数。在决策阶段, 支持向量机以式(2)判断样本位于最优分界面哪一侧。

$$D(x) = w^T x + b \quad (2)$$

Joachims 最早将 SVM 方法应用到网页分类, 并通过组合核函数的方法把支持向量机用于网页分类<sup>[3]</sup>。Yu, Han 等人<sup>[4]</sup>提出了一种仅考虑正例网页样本的分类方法: PEBL。

## 1.2 FTSM 的基本原理

直推式支持向量机(TSM)<sup>[5]</sup>是直推式学习理论与支持向量机的结合, 适用于训练集和测试集分布有一定差别的样

本。刘宏等在 TSM 基础上提出了 FTSM。FTSM 最大的特点是构建了新的目标函数。FTSM 的目标函数如式(3):

$$\frac{1}{2} w^T w + C \sum_i \xi_i + C^* \sum_j u_j \xi_j^* \quad (3)$$

$$\text{st. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0;$$

$$y_j(w^T x_j + b) \geq 1 - \xi_j^*, \quad \xi_j^* \geq 0$$

其中,  $w$  是最优分界面的方向向量,  $b$  是阈值,  $x$  是样本向量,  $y$  是类别标记,  $\xi_i$  和  $\xi_j^*$  是训练集和测试集的松弛变量;  $C$  和  $C^*$  是训练集和测试集的影响参数。 $\mu$  是 FTSM 增加的模糊隶属度因子, 它通过分析网页与网页间的链接关系得到。

在 FTSM 中, 全部训练集正例样本和反例样本被分别集中成两个样本点:  $p^+$  和  $p$ 。测试集样本对于正例或反例的模糊隶属度因子  $u^+$  或  $u$  根据该样本与  $p^+$  或  $p$  的链接关系来判断。 $u^+$  和  $u$  3 个相似系数加权而成。

## 2 基于潜在链接分析的 FTSM

## 2.1 FTSM 的不足

FTSM 有两处值得改进。首先, FTSM 中计算模糊隶属度因子的方法需要解有向图中任意两点之间的最短路径问题, 是一个复杂度相当高的算法。此外, 3 个相似系数之间权重凭经验确定, 随意性大。

其次, FTSM 忽略了网页本身对于分类的贡献。实际上, 网页可分为“导航页面”和“实体页面”两类。如果不加区分地训练这两类页面, 可能引起偏差。

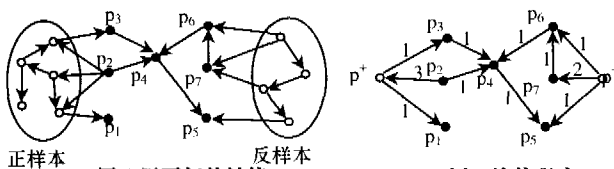
## 2.2 潜在链接分析

潜在语义分析(LSA)<sup>[6]</sup>的基本原理是对“共现表”进行奇异值分解, 并选取最大的若干个奇异值重构共现表, 重构后

**作者简介:** 王 晔(1975—), 男, 博士生, 主研方向: 机器学习, 数据挖掘; 黄上腾, 教授

**收稿日期:** 2005-07-24 **E-mail:** wangye@sju.edu.cn

的共现表反映了特征之间的潜在共现信息，我们将这一过程称作“潜在链接分析”(LLA)。下面以图1为例，详细说明潜在链接分析。



在图1中，左侧白点代表正训练样本，右侧白点代表反训练样本，黑点代表测试样本，图中的一条有向边代表网页之间存在超链接。合并正例样本成  $p^+$ ，合并反例样本成  $p^-$ ；并定义样本  $p_i$  到  $p_j$  的“连接强度”为由  $p_i$  指向  $p_j$  的有向边总数，如图2所示。初始连接强度矩阵  $A$  包含了任意两样本之间的连接强度，如表1。

表1 初始连接强度矩阵  $A$

	$p^+$	$p^-$	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$	$p_7$
$p^+$	×	0	1	0	1	0	0	0	0
$p^-$	0	×	0	0	0	0	0	1	2
$p_1$	0	0	×	0	0	0	0	0	0
$p_2$	3	0	0	×	0	1	0	0	0
$p_3$	0	0	0	0	×	1	0	0	0
$p_4$	0	0	0	0	0	×	1	0	0
$p_5$	0	0	0	0	0	0	×	0	0
$p_6$	0	0	0	0	0	1	0	×	0
$p_7$	0	0	0	0	0	0	1	1	×

表1中， $p_i$ 到其它网页的连接强度位于  $p_i$ 所在的行；其它网页到  $p_i$ 的连接强度位于  $p_i$ 所在的列。表1中的“×”表示网页与自身的连接强度，我们以最大的不同样本间的连接强度代替它。对表1进行奇异值分解，取最大的两个奇异值重构邻接矩阵，得到矩阵  $A'$ ，列于表2。

表2 LLA 后的连接强度矩阵  $A'$

	$p^+$	$p^-$	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$	$p_7$
$p^+$	2.12	-0.14	0.44	1.27	0.57	<b>0.82</b>	0.07	-0.02	-0.20
$p^-$	-0.27	1.34	-0.09	-0.12	0.04	<b>0.57</b>	0.87	1.56	2.08
$p_1$	0.38	-0.04	0.08	0.23	0.10	0.14	0.00	-0.02	-0.06
$p_2$	3.16	-0.12	0.65	1.89	0.86	1.27	0.17	0.08	-0.15
$p_3$	0.73	0.08	0.15	0.44	0.21	0.35	0.11	0.15	0.14
$p_4$	<b>0.72</b>	<b>0.34</b>	0.14	0.44	0.23	0.48	0.28	0.46	0.54
$p_5$	0.06	0.36	0.00	0.04	0.04	0.20	0.24	0.42	0.56
$p_6$	2.12	-0.14	0.44	1.27	0.57	0.82	0.07	-0.02	-0.20
$p_7$	-0.27	1.34	-0.09	-0.12	0.04	0.57	0.87	1.56	2.08

表2反映了任意两样本之间的潜在连接强度，例如  $p_4$ 不和  $p^+$ 以及  $p^-$ 直接相连，但在表2中， $p^+ \sim p_4$  ( $p_4 \sim p^+$ )以及  $p^- \sim p_4$  ( $p_4 \sim p^-$ )的潜在连接强度均不为0，如黑体数字所示。 $A'$ 中各值还应经过下式进行归一化：

$$na_{ij} = \frac{a'_{ij} - \min(a'_{ij})}{\max(a'_{ij}) - \min(a'_{ij})}, \quad \forall a'_{ij} \in A' \quad (4)$$

取测试样本  $p_i$  在  $p^+$ 和  $p^-$ 所在的行和列中的元素平均值作为测试样本  $p_i$  与  $p^+$ 和  $p^-$ 的归一化的潜在连接强度，即

$$u_{i,c}^m = (na_{mi} + na_{im}) / 2, \quad m \in \{+, -\} \quad (5)$$

### 2.3 模糊隶属度函数

我们利用“内容链接比”来反映网页自身的分类权重，“内容链接比”(CLR)是指一个页面的大小(以千字节计)与页面中的超链接数之比。我们认为页面  $p_i$  的 CLR 小于0.5时，该页面具有一定的“导航”性质，其权重应予折减。网页的权重系数  $\lambda_{i,imp}$  定义为

$$\lambda_{i,imp} = \begin{cases} 2 \times CLR_i, & 0 \leq CLR_i \leq 0.5 \\ 1, & CLR_i > 0.5 \end{cases} \quad (6)$$

测试样本  $p_i$  模糊隶属度函数被定义为  $p_i$  与  $p^+$ 和  $p^-$ 的归一化的潜在连接强度和网页权重系数的乘积，即

$$u_i^m = u_{i,c}^m \times \lambda_{i,imp}, \quad m \in \{+, -\} \quad (7)$$

根据式(4)~式(6)可知， $u_i^m$ 的值介于0和1之间，符合模糊隶属度函数的要求。本文提出的模糊隶属度函数计算方法的主要步骤是奇异值分解，它的复杂度是  $O(k^3 N^2)$ ，其中  $k$  是保留特征的个数。由于 SVM 中的样本仅有正、反两类，因此潜在链接分析时保留3~4个特征即可。这样，其复杂度降为  $O(N^2)$ 。

### 2.4 改进的目标函数

根据上面的分析，构造 LLA\_FTSVM 的目标函数：

$$\begin{aligned} \min & \frac{1}{2} w^T w + C \sum_i u_i \xi_i + C^* \sum_i u_i^m \xi_i^* \\ \text{st. } & y_i (w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0; \\ & y_j (w^T x_j + b) \geq 1 - \xi_j^*, \quad \xi_j^* \geq 0 \end{aligned} \quad (8)$$

其中， $u_i^m$ 是测试样本的模糊隶属度因子。与 FTSVM 相比，LLA\_FTSVM 的目标函数的最大不同是为训练集样本的惩罚项也增加了一个模糊隶属度因子  $u_i$ 。令  $u_i$  等于网页的分类重要性系数  $\lambda_{i,imp}$ ，其作用是减小训练集中“导航页面”的影响。

LLA\_FTSVM 的学习过程可以被看作是以  $(w, b, \xi_1, \xi_2, \dots, \xi_m, \xi_1^*, \xi_2^*, \dots, \xi_l^*)$  为参数最小化目标函数式(8)，其中  $n$  是训练集样本总数， $l$  是测试集样本总数。通过二次规划法解上述问题，将其转变为以拉格朗日乘数为参数的最小化问题，即

$$\begin{aligned} \min_{\alpha} & \frac{1}{2} \sum_{i,j=1}^{n+l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n+l} \alpha_i \\ \text{st. } & \sum_{i=1}^{n+l} \alpha_i y_i = 0; \quad \forall y_i \in TRS: 0 \leq \alpha_i \leq C u_i; \\ & \forall y_j \in TES: 0 \leq \alpha_j \leq C^* u_j^m \end{aligned} \quad (9)$$

其中， $TRS$  是训练集； $TES$  是测试集。为了对式(8)进行求解，LLA\_FTSVM 在 FTSVM 的基础上改进了训练算法，LLA\_FTSVM 的训练算法如下：

(1) 预处理网页样本，并计算各样本的模糊隶属度因子。指定参数较大的  $C$  和较小的  $C^*$ ，用训练样本训练初始支持向量机，并按训练集样本比例指定测试样本中正标签样本数  $N_p$ 。

(2) 用步骤(1)训练的支持向量机对测试样本进行分类，将输出值最大的  $N_p$  个测试样本暂赋予正标签，其余的暂赋予反标签，并根据  $N_p$  和  $C^*$  计算正反例的影响参数  $C_+^*$  和  $C_-^*$ 。

(3) 按式(8)的目标函数重新训练所有样本，对于测试样本，反复执行(内循环)：寻找一对不同类别的样本，它们的松弛变量满足均大于0，且和大于2；如找到，则交换它们的标签值并重新训练，直到所有测试样本均不符合上述条件为止。

(4) 同比例增大影响参数  $C_+^*$  和  $C_-^*$ ，若  $C_+^*$  和  $C_-^*$  均小于  $C^*$ ，则返回执行步骤(3)(外循环)；否则算法结束并输出测试集样本的标签。

## 3 试验结果

为检验本文提出的新方法在网页分类领域的性能，我们进行了 TSVM, FTSVM 以及 LLA\_FTSVM 三者的对比试验。

### 3.1 试验设置

试验在网页基准测试集 WebKB 上进行。WebKB 包含了从美国4所大学的计算机系网站中选取的8282个页面。这些页面被分为7个类别。

我们试验的任务是从 WebKB 测试集中识别出学生、课程、教员类别的网页。在全部 WebKB 页面中随机选择  $n$  个

页面作为训练集,其中属于类 A 的样本作为正例,其他作为反例。再从全部页面中随机地选择 2 000 个页面作为测试集,学习机的任务就是从测试集中判断出属于类 A 的样本。

我们用 VSM 来表示网页中的纯文本。网页样本经过删除停用词、取词干后整理成高维词空间的向量。测试网页之间的超链接按第 2.2 节所述整理成连接强度矩阵的形式。各试验均在 Matlab 环境下实现。

### 3.2 试验结果

我们以能够综合反映分类准确度 P 和召回度 R 的  $F_1$  指标来衡量 TSVM, FTSVM 以及 LLA\_FTSVM 的性能。P, R 以及  $F_1$  的计算公式如下:

$$P = \text{标记正确的正样本数} / \text{标记为正样本的样本数}$$

$$R = \text{标记正确的正样本数} / \text{实际的正样本数}$$

$$F_1 = 2PR / (P+R)$$

以  $F_1$  为衡量指标,应用 TSVM, FTSVM 以及 LLA\_FTSVM 对学生、课程、教员 3 个类别,在不同的训练样本数  $n$  的条件下,进行识别试验的结果见表 3、表 4。

表 3 识别结果 (左:  $n = 100$ , 右:  $n = 200$ )

类别	TSVM	FTSVM	LLA	类别	TSVM	FTSVM	LLA
学生	81.5%	83.5%	90.3%	学生	84.6%	89.6%	92.3%
课程	90.1%	89.7%	91.8%	课程	89.6%	90.3%	92.4%
教员	56.4%	67.5%	68.2%	教员	61.4%	70.5%	71.2%

表 4 识别结果 (左:  $n = 400$ , 右:  $n = 800$ )

类别	TSVM	FTSVM	LLA	类别	TSVM	FTSVM	LLA
学生	86.7%	90.0%	93.9%	学生	89.6%	90.5%	93.9%
课程	88.7%	88.3%	92.4%	课程	90.5%	90.6%	92.4%
教员	65.4%	77.1%	78.8%	教员	75.2%	78.8%	81.4%

我们还对 FTSVM 和 LLA\_FTSVM 计算模糊隶属度所用的时间进行了比较,比较结果见表 5。表中单位为 s。

表 5 超链接分析时间的比较

类别	FTSVM	LLA_FTSVM
学生	927.1	94.9
课程	1 370.3	85.6
教员	1 628.4	89.7

(上接第 11 页)

10%的另一主导流量,两个流量的合成导致了 LIDC 框架下的  $n(a)$  的估计分段点从 8 变为 5。但这需要进一步的分析和验证。

### 4 结论

上节的分析讨论表明,LIDC 在分析网络流量的多尺度行为方面非常有效。不仅在全尺度范围内能刻画网络流量的尺度特性,而且能确定出某一尺度下的有效分析范围。通过使用该分析框架,还得到一些结论,即网络流量中的尺度分段点具有不确定性,并不只是落在倍频  $j=8$  ( $=2.56s$ )<sup>[5]</sup>上,从得到的数据流还可能落在倍频  $j=5$  处;受到特定网络蠕虫影响的数据流甚至出现了保持幂律关系、单一尺度以及尺度不变性等特点。从讨论中可知,数据流出现的与已有研究不同的特性是由于数据流中协议组成及其百分比的不同引起的,每种应用协议自己特定的协议机制都对数据流的最终形态产生了特定的效果。

### 3.3 讨论

通过观察表 3、表 4,发现 LLA\_FTSVM 在总体上性能比 FTSVM 和 TSVM 要好。特别是“课程”类别。LLA\_FTSVM 的识别结果有了明显的提高。我们认为,同时考虑网页的分类重要性信息可以明显提高分类的性能。

通过观察表 5,我们发现 LLA\_FTSVM 的超链接分析时间大致相等,且显著小于 FTSVM 的超链接分析时间。

### 4 结论

本文是在 FTSVM 的基础上进一步深入研究,并提出了一种基于潜在链接分析、并结合网页权重的 LLA\_FTSVM 网页分类方法。

在 WebKB 网页基准测试数据上的试验结果表明,本文提出的 LLA\_TSVM 在大大缩短超链接分析时间的基础上,获得了比 TSVM 和 FTSVM 更好的准确度和召回度,在网页分类方面取得了令人满意的效果。

### 参考文献

- 1 Liu H, Huang S T. Fuzzy Transductive Support Vector Machines for Hypertext Classification[J]. International Journal of Uncertainty, Fuzziness and Knowledge Based System, 2004; 12(1).
- 2 Vapnik V. Statistical Learning Theory[M]. New York: Wiley, 1998.
- 3 Joachims T, Cristianini N, Shawe-Taylor J. Composite Kernels for Hypertext Categorisation[C]. Proceedings of the International Conference on Machine Learning, 2001: 250-257.
- 4 Yu H, Han J, Chang K C. PEBL: Positive Example Based Learning for Web Page Classification Using SVM[C]. SIGKDD '02 Edmonton, Canada, 2002.
- 5 Joachims T. Transductive Inference for Text Classification Using Support Vector Machines [C]. Proceedings of 16<sup>th</sup> International Conference on Machine Learning, 1999.
- 6 Scott D, Dumais S T. Indexing by Latent Semantic Analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

### 参考文献

- 1 程 华, 邵志清, 房一泉. Internet 流量的多重分形分析 [J]. 通信学报, 2005, 26(1A):27-30.
- 2 Arneodo A, Muzy J F, Roux S G. Experimental Analysis of Self-similarity and Random Cascade Processes: Application of Fully Developed Turbulence Data[J]. Journal de Physique II (France), 1997, 7(2): 363-370.
- 3 Arneodo A, Bacry E, Muzy J F. Random Cascades on Wavelet Dyadic Trees [J]. Journal of Mathematical Physics, 1998, 39 (8).
- 4 Chainais P, Abry P, Pinton J. Intermittency and Coherent Structures in a Swirling Flow: A Wavelet Analysis of Joint Pressure and Velocity Measurements[J]. Physics of Fluids, 1999, 11(11): 3524-3539.
- 5 Veitch D, Abry P, Flandrin P, et al. Infinitely Divisible Cascade Analysis of Network Traffic Data [C]. Proc. ICASSP' Conference, Turkey, 2000.