

面向大规模信息的用户分类方法研究

刘忠宝¹, 赵文娟²

(1 中北大学 电子与计算机科学技术学院, 山西 太原 030051;

2 山西大学商务学院 信息学院, 山西 太原 030031)

摘 要: 信息主动推送是个性化搜索引擎的发展方向, 如何通过访问记录获取用户兴趣以及如何将兴趣相同或相近的用户归为一类, 是个性化信息主动推送技术面临的主要问题. 支持向量机是一种广泛使用的分类方法, 但在互联网环境下, 其面临大规模数据分类和隐私泄露问题. 为了进一步提高分类效率, 提出面向大规模信息的用户分类方法. 该方法不仅解决了大规模用户分类问题, 还保证分类过程中用户信息安全.

关键词: 用户分类; 支持向量机; 隐私保护; 大规模信息

中图分类号: TP392

文献标识码: A

文章编号: 1000-7180(2013)06-0038-05

Research on User Classification Method Based on Large-scale Information

LIU Zhong-bao¹, ZHAO Wen-juan²

(1 School of Electronics and Computer Science Technology, North University of China, Taiyuan 030051, China;

2 School of Information, Business College of Shanxi University, Taiyuan 030031, China)

Abstract: Information active push is the trend of personalized search engine. How to get user interests and classify the users of the same interest are important. Support Vector Machine (SVM) is widely used, while it faces some challenges, including classification of large-scale datasets and information leakage. In order to solve the above problems, User Classification Method based on Large-scale Information (UCM) is proposed in this paper, it can effectively solve the large-scale information classification and preserve user information during classification.

Key words: user classification; SVM; private preserving; large-scale information

1 引言

搜索引擎是查找信息的主要工具,但其查询返回结果相关率并不理想. 为了进一步提高信息服务的质量和用户满意度,开发性能更加优越的个性搜索引擎成了当务之急^[1]. 个性化搜索引擎通过收集和分析用户的浏览信息来获取用户的行为和兴趣,从而达到返回用户个性化搜索结果的目的. 所谓个性化是指通过分析用户输入的检索要求、用户浏览的页面以及用户的使用日志等获取满足用户个性化需求信息^[2]. 目前,个性化搜索引擎大多是针对用户检索要求“被动地”提供信息服务. 未来信息服务

发展方向是信息主动推送,传统的个性化搜索引擎关注的是用户的差异性,而信息主动推送更强调用户兴趣的相关性. 因此,如何通过访问记录获取用户兴趣以及如何将兴趣相同或相近的用户归为一类,是个性化信息主动推送面临的主要问题. 此外,信息主动推送还受制于互联网海量信息处理技术的发展.

支持向量机 (Support Vector Machine, SVM)^[3]是当前广泛使用的分类方法之一,但在网络环境下,其面临两大问题:(1)大规模数据分类问题;(2)隐私泄露问题. 因此,本文在分析 SVM 的基础上,提出面向大规模信息的用户分类方法 (User

收稿日期: 2012-08-13; 修回日期: 2012-10-27

基金项目: 山西省科技厅软科学项目(2010041013-06); 山西大学商务学院科研基金项目(XS2011005)

Classification Method based on Large-scale Information, UCM). 该方法不仅有效地解决了 SVM 面临的两大问题,而且适用于解决互联网用户分类问题。

2 支持向量机及其面临的主要问题

支持向量机 SVM 是一种基于统计学习理论的机器学习方法。该最优化问题可描述为

线性形式:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (1)$$

$$\text{s. t. } y_i(w^T x_i + b) \geq 1 - \xi_i \quad i=1, \dots, N, \xi_i \geq 0 \quad i=1, \dots, N \quad (2)$$

式中, $x_i \in R^d$ ($1 \leq i \leq N$) 为输入数据, $y_i \in \{1, -1\}$ 为类别标签。C 为惩罚因子, 它控制对错分信息的惩罚程度, $C=0$ 时表示线性可分, $C>0$ 时表示线性不可分; 对于线性不可分或事先未知是否线性可分的情况, 通过引入松弛因子 ξ_i 允许错分数据的存在。

其决策函数为

$$f(x) = \text{sgn}(w^T x + b^*) \quad (3)$$

式中, $w = \sum_{i=1}^N \alpha_i^* y_i x_i$, $b^* = y_j - \sum_{i=1}^N y_i \alpha_i^* x_i^T x_j$ 且 $\alpha^* > 0$ 。

核化形式:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

$$\text{s. t. } y_i(w^T \varphi(x_i) + b) \geq 1 - \xi_i \quad i=1, \dots, N, \xi_i \geq 0 \quad i=1, \dots, N \quad (4)$$

式中, $\varphi(x_i)$ 表示从原始空间到高维特征空间的映射。

其决策函数为

$$f(x) = \text{sgn}\left(\sum_{i=1}^N \alpha_i^* y_i k(x_i, x) + b^*\right) \quad (5)$$

式中, $b^* = y_j - \sum_{i=1}^N y_i \alpha_i^* k(x_i, x_j)$ 且 $\alpha^* > 0$ 。

由上述分析可知: 线性 SVM 本身实现了隐私保护, 原因是线性 SVM 决策函数 $f(x) = w^T x + b$ 中权重 w 是所有支持向量的线性组合, 每个支持向量的敏感内容由于权重 w 的存在而不被暴露。但线性 SVM 使用范围有限, 核方法的引入导致线性 SVM 隐私保护作用消失: 式(5)表明判断样本 x 的类属必须事先得知所有训练样本信息, 这对隐私保护带来巨大挑战。此外, 当面临大规模数据时, SVM 的分类性能受到很大影响。

3 面向大规模信息的用户分类方法

为了能够解决 SVM 面临的两大问题, 本文提出面向大规模信息的用户分类方法 UCM。该方法首先对页面进行预处理, 得到页面集合; 然后将页面集合划分成若干子集并在各子集上运行分类超平面 (Separating Hyperplane, SH) 算法得到相应的决策函数 $f_i(x)$; 最后通过非线性函数将上述决策函数 $f_i(x)$ 集成, 得到最终的决策函数 $f(x)$, 进而实现用户分类。为了表述方便, 后两步称为基于分类超平面的非线性集成分类方法 (Nonlinearly Assembling Classification Method based on Separating Hyperplane, NACM)。UCM 的流程图如图 1 所示。

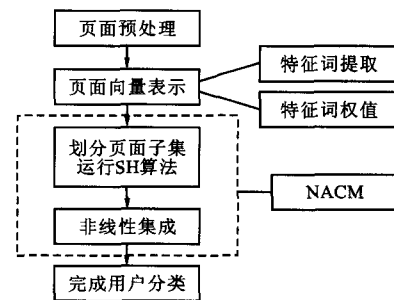


图1 UCM流程图

3.1 页面的表示

通过挖掘用户访问的页面信息可以发现用户的兴趣, 从而为将兴趣相同或相近的用户归为一类奠定基础。因此, 首先应将页面进行预处理及向量表示。

(1) 页面预处理

页面预处理主要包括页面清洗, 即清除掉一些与研究无关的文件, 如图片文件及脚本程序等。

(2) 页面的向量表示

面对杂乱无章的页面首先应对其进行特征表示, 然后根据页面特征所反映出来的内容得到页面的向量表示。目前比较成熟的页面特征表示方法有布尔逻辑模型^[4]、向量空间模型 (Vector Space Model, VSM)^[5]、概率模型^[6]等。本文页面表示方法基于 VSM。

本文对关键词的考察重点在于对其权值的考察, 因此页面特征表示采用加权特征词方法, 该方法包含两步: ① 页面特征词提取; ② 特征词权值计算。

页面特征词提取方法如下:

Step1: 将页面转化成文本并保留某些重要标记信息;

- Step2:对文本文件切词处理^[7];
- Step3:去掉与页面内容无关的虚词;
- Step4:去掉低频词,低频阈值^[8]确定方法见表 1;
- Step5:剩下的词作为特征词,并保留其出现频率.

表 1 低频阈值确定方法

文章长度	低频阈值
(0,200]	2
(200,4 000]	3
(4 000,10 000]	4
(10 000,25 000]	5
(25 000,+∞)	6

页面特征词权值计算方法如下:

Step1:根据特征词 t_i 在页面中出现的位置和次数 f_i 计算其频率:

$$f_i' = f_i \times s_i \quad (i=1,2,\cdots,N) \tag{6}$$

式中, s_i 为特征词 t_i 对应页面标记的权系数. s_i 的取值见表 2.

Step2:由于页面长短不一,式(7)实现规范化处理:

$$f_i = \frac{f_i'}{\sqrt{\sum_{i=1}^N f_i'^2}} \quad (i=1,2,\cdots,N) \tag{7}$$

表 2 HTML 部分标记权重设置表

HTML 标记	权重
<TITLE>	1
<H1> <H2> <H3>	0.8
 	0.7
<BODY>	0.5

在得到页面特征词及其权值后,可将页面表示为 $p=\{(k_1,tf_1),(k_2,tf_2),\cdots,(k_N,tf_N)\}$ ($i=1,2,\cdots,N$),其中 k_i 和 tf_i 分别为页面特征词及其权值.

3.2 基于分类超平面的非线性集成分类方法

3.2.1 分类超平面

针对 SVM 面临的隐私泄露问题,文献[9]中提出分类超平面(Separating Hyperplane, SH):

$$\begin{aligned} d(x) &= \text{sgn}(y_i x_i^T x - y_j x_j^T x - b) \\ &= \text{sgn}(w^T x - b) \end{aligned} \tag{8}$$

式中, $w = y_i x_i - y_j x_j$.

该超平面工作原理如下:对于两类数据,两类中各选一个数据点,将两点连线的法平面作为候选分

类面. 由于存在无穷多个这样的分类面,因此规定该分类面至少经过一个数据点. 选取误分率最低的分类面作为最优分类面. 不难看出,SH 的时间复杂度为 $O(n^3)$. 上述思想如图 2 所示,其中 $\circ$ 和 $+$ 分别表示两类数据.

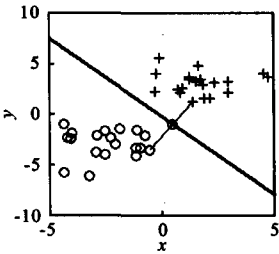


图 2 分类超平面工作原理图

与核 SVM 相比,SH 的最大优势在于仅需要三个数据点就能确定分类面. 这在很大程度上降低了隐私泄露的风险. 然而随着数据规模不断增大,SH 的时间规模急剧增加. 由此可见,SH 仍无法解决大规模数据分类问题. 基于上述分析并借鉴管理学中的协同思想,本文提出基于分类超平面的非线性集成分类方法(Nonlinearly Assembling Classification Method based on Separating Hyperplane, NACM).

3.2.2 NACM

在管理学中,协同指协调两个或者两个以上的不同资源或者个体,协同一致地完成某一目标的过程或能力. 在一个系统内,若系统中各子系统能很好配合,多种力量就能集聚成一个总力量,形成大大超越原各自功能总和的新功能. 将上述思想运用于大规模页面分类,可得下面 NACM 方法:

- Step1:将页面集合 D 分为 M 个子集 $\{D_1, D_2, \cdots, D_M\}$ 并在每个数据子集 D_i ($i=1,2,\cdots,M$) 上运行 SH 算法得到相应的决策函数 $f_i(x)$;
- Step2:通过非线性函数将上述决策函数 $f_i(x)$ 集成,得到最终的决策函数 $f(x)$.

由 NACM 方法可以看出:NACM 不仅最大限度地发挥 SH 具有的优势,而且将 SH 的适用范围从小规模数据扩展到中大规模数据;NACM 通过非线性函数集成的方式将 SH 从原始线性空间推广到非线性空间,有效地扩大了其适用范围.

3.2.3 页面集合划分方法

为了避免引入“数据不平衡”问题,本文采用“随机等分法”对数据集进行划分. 假设两类页面数分别为 n^+ 和 n^- ,则两类页面数之比 $p = n^+ / n^-$. 若将两类页面组成的集合 T 等分成 M 份,则每个子集

的规模为 $s = \lceil n^+ + n^- / M \rceil$, 其中 $\lceil \cdot \rceil$ 表示取整. 随机等分法要求每个子集中的页面从集合 T 中随机选取, 各子集满足两类页面数之比及子集规模分别等于 p 和 s 且各子集无交集. 这样, 在一定程度上避免数据子集出现只有一类数据或一类数据过多而另一类过少等“数据不平衡”问题.

3.2.4 非线性集成方法

NACM 试图通过页面集合划分以及非线性集成两步解决大规模页面分类问题. 其中, 非线性集成函数的选取至关重要.

当前常用的核函数有: Gaussian 核函数、多项式核函数、Sigmoid 核函数、Epanechnikov 核函数等. 其中径向基函数(Radial Based Function, RBF)具有若干优良性质, 在实际应用中特别在非线性分类方面表现出一定优势. 因此, 本文将 RBF 核函数的非线性特性和 SH 的隐私保护特性有机地结合起来, 从而将 SH 的适用范围从线性空间推广到非线性空间. 同时, 为了将各页面子集上的分类结果有效集成, 因此引入非线性输出权重 α_i ($i = 1, 2, \dots, M$), 其满足 $\alpha_i \geq 0$. 综上, 非线性集成函数可表示为

$$f(x) = \sum_{i=1}^M \alpha_i \exp(-\|x - \mu_i\|^2 / h)(w_i^T x - b_i)$$

(9)

其中径向基核函数中的参数 μ_i 通过求解各数据子集的中心获得; 参数 h 通过网格搜索获得. 由“数据集划分方法”可知: 各页面子集的分布性状近似, 因此特别地令 $\alpha_i = 1/M$.

3.3 重要性质

性质 1: NACM 的 VC 维小于等于 9.

证明: (1) 由 VC 维定义可知: 对于一个函数集, 如果存在 h 个样本能被函数集中的函数按所有可能的 2^h 种形式分成两类, 则称函数集能把 h 个样本打散. 函数集的 VC 维就是它能打散的最大页面数 h .

(2) 由 SH 算法可知: 两类页面共产生有序页面对的数量为 $h(h-1)$; 由于分类面至少经过一个页面向量, 则共有 $h^2(h-1)$ 种形式将页面集合分开.

当 $h > 9$ 时, 有 $2^h > h^2(h-1)$. 即当可打散的最大页面数大于 9 时, 式(8)表示的超平面类不能打散它们. 由此可知: NACM 的 VC 维小于等于 9.

由 VC 维理论可知: VC 维反映分类器的学习能力, VC 维越大, 分类器的性能越差; 反之, 分类器的性能越好. 因此, 性质 1 从理论上保证 NACM 对大规模页面分类的有效性.

4 实验

为了检验面向大规模信息用户分类方法 NACM 的有效性, 通过模拟实验进行验证. 实验选取 2012 年 3 月 1 日至 2012 年 3 月 30 日 30 天内 8 个用户访问的网页作为实验对象. 用户及其浏览的页面数如表 3 所示.

表 3 用户及其浏览页面数

用户 ID	页面集合 ID	浏览页面数
User_1	S_1	3 218
User_2	S_2	2 854
User_3	S_3	3 102
User_4	S_4	4 533
User_5	S_5	2 345
User_6	S_6	1 000
User_7	S_7	2 817
User_8	S_8	3 211

对上述页面进行向量表示后, 为了将用户全部归类, 故将上述页面集合两两组合并分别运行 SVM、SH 和 NACM. 随机选取页面集合的 70% 作为训练样本, 余下的 30% 作为测试样本. 将页面集合近似等分为 M 份, 并令 $M = 20$. 实验结果如表 4 所示, 其中 C_1 、 C_2 、 C_3 分别表示各用户的类属; “—”表示在有限时空范围内无法求解. 上述算法的分类时间如表 5 所示.

表 4 对比实验结果

用户 ID	SVM	SH	NACM
User_1	—	—	C_1
User_2	—	—	C_1
User_3	—	—	C_3
User_4	—	—	C_2
User_5	—	—	C_2
User_6	C_1	—	C_1
User_7	—	—	C_2
User_8	—	—	C_3

表 5 算法分类时间 s

页面集合	SVM	SH	NACM
s_1	—	—	9.8
s_2	—	—	9.3
s_3	—	—	9.5
s_4	—	—	13.4
s_5	—	—	6.9
s_6	1 200.3	—	3.0
s_7	—	—	5.1
s_8	—	—	8.8

由表 4 和表 5 可以看出:在面临较大规模页面信息时,传统方法 SVM 和 SH 基本上无法求解(除 S_6 外),而 NACM 能快速得到分类结果.由此可见,NACM 在解决大规模页面分类问题上的有效性是传统方法所不具备的.而且 NACM 在一定程度上对用户信息进行了隐私保护.

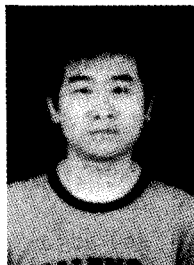
5 结束语

如何从用户访问的页面信息对用户进行归类是进行信息主动推送的关键一步.本文在分析当前主流分类方法 SVM 的基础上,提出面向大规模数据的用户分类方法 UCM.该方法不仅解决了大规模用户分类问题,还保证分类过程中用户信息安全.上述优势使得 UCM 非常适合在互联网环境下对海量用户信息分类,也适用于其他海量数据分类问题.

参考文献:

- [1] 徐谦.网络信息检索的智能化趋势[J].图书馆理论与实践,2006(2):63—65.
- [2] 曹红兵.搜索引擎的个性化检索研究[J].图书情报工作,2007,51(3):139—136.
- [3] Pal M, Foody G M. Feature selection for classification of hyper spectral data by SVM [J]. IEEE Trans. on Geoscience and Remote Sensing, 2010, 48 (5): 2297—2307.
- [4] 刘红泉,张亮峰.布尔逻辑检索模型的分析探讨[J].现代情报,2004(9):4—6.
- [5] 唐明伟,卞艺杰,陶飞飞.基于语义向量空间模型的文档检索系统研究[J].情报杂志,2010(5):167—170.
- [6] 刘华.文本分类相似度模型和概率模型的实现与比较[J].现代图书情报技术,2006(4):53—55.
- [7] 许林杰.中文文本分词研究[D].山东:山东师范大学,2003.
- [8] 顾立帆,王永成.联想树分析法及其在无词库中文自动标引中的应用[J].情报学报,1992,11(5):354—360.
- [9] Marcin O. New separating hyperplane method with application to the optimization of direct marketing campaigns[J]. Pattern Recognition Letters, 2011(32): 540—545.

作者简介:



刘忠宝 男,博士.研究方向为机器学习与信息检索.

(上接第 37 页)

- [5] Steve S Liao, P H W, Hong Wang, et al. Post-pass binary adaptation for software-based speculative pre-computation[C]//Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2002). Germany: Berlin, Association for Computing Machinery, 2002: 117—128.
- [6] Grunwald D, D J a. Prefetching using markov predictors [J]. IEEE Transactions on Computers, 1999, 48 (2): 121—133.
- [7] Nesbit KJ, Smith J E. Data cache prefetching using a global history buffer [J]. IEEE Micro, 2005, 25(1): 90—97.
- [8] Diaz P, Cintra M. Stream chaining: exploiting multiple levels of correlation in data prefetching[C]//Proceedings of the 36th International Symposium on Computer Architecture. USA: Austion, Texas, IEEE Comput. Soc, 2009: 81—92.
- [9] Stephen Somogyi, T F W, Anastassia Ailamaki, et al. Spatial memory streaming[C]//Proceedings of the 33rd International Symposium on Computer Architecture. USA: Boston, IEEE Comput. Soc, 2006: 252—263.

作者简介:



朱会东 (1975—)硕士,讲师.研究方向为高性能计算、缓存性能优化.

段赵磊 男,博士,讲师.