基于 SVM 的不良文本信息识别①

吕洪艳, 杜 鹃

(东北石油大学 计算机与信息技术学院, 大庆 163318)

摘 要: 不良文本识别的实际应用中, 大多数文本之间总有交界甚至彼此掺杂, 这种非线性不可分问题给不良文本识别带来了难度. 应用 SVM 通过非线性变换可以使原空间转化为某个高维空间中的线性问题, 而选择合适的核函数是 SVM 的关键. 由于单核无法兼顾对独立的不良词汇和词汇组合的识别, 使识别准确率不高, 而且也无法兼顾召回率. 针对不良文本识别的特定应用, 依据 Mercer 定理结合线性核与多项式核提出了一种新的组合核函数, 这种组合核函数能兼顾线性核与多项式核的优势, 能够实现对独立的不良词汇以及词汇组合进行识别. 在仿真实验中评估了线性核、齐次多项式核以及组合核函数, 实验结果表明组合核函数的识别准确率与召回率都比较理想.

关键词: SVM; 组合核函数; 不良文本; 信息识别; 召回率

Undesirable Text Recognition Based on SVM

LV Hong-Yan, DU Juan

(Institute of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: In practical application of undesirable text information identification, most of the text always have intersection even doped with each other. The nonlinear non-separable problem has brought difficulty to undesirable text information identification. SVM can make a nonlinear problem in the original space into a linear problem in high dimension space by nonlinear transformation. And the key of the SVM is to choose the appropriate kernel function. A single kernel function can not recognize the independent undesirable vocabulary and vocabulary combination at the same time, so the recognition accuracy rate is not high and the Rcall value is not ideal. For the specific application of undesirable text information identification, combining with linear kernel and homogeneous polynomial kernel it structured a new combination kernel function according to the Mercer theorem. This combination kernel function has the advantage of both linear kernel and polynomial kernel, and could identify the independent undesirable vocabulary and vocabulary combination. Then it evaluated the linear kernel, homogeneous polynomial kernel and combination kernel function in the sample experiment. The experimental results showed that the recognition accuracy rate and the Rcall value of combination kernel function was more ideal than other kernel functions.

Key words: SVM; combination kernel function; undesirable text; information identification; recall

1 引言

随着信息技术的快速发展, 网络信息安全问题日益突出, 特别是网上色情、暴力等不良信息的泛滥, 已经严重影响了青少年的健康成长, 如何对这些不良信息进行过滤, 保证青少年健康上网已成为整个社会亟待解决的问题[1]. 目前国内外针对网页不良信息的过滤技术主要有以下四种: 基于因特网内容分级平台(PICS)过滤、数据库过滤、关键词过滤以及基于内容理解的过滤[2,3]. 由于在实际应用中, 不良信息不会按照 PICS 的标准贴标签, 所以基于 PICS 过滤的效果受到限制. 数据库过滤对一些更改 IP、URL 地址或采用多级代理方式的网页信息无法过滤. 目前来说基于关

① 收稿时间:2014-10-12;收到修改稿时间:2014-11-28

键词的过滤速度较快, 但由于不考虑信息内容的关联性, 错报率较高. 基于内容理解的过滤技术能动态地辨别文档内容的实际含义, 识别准确率较高, 得到了广泛的应用. 基于文本内容理解的信息过滤模型主要有向量空间模型(VSM)、贝叶斯决策模型、神经网络模型、潜在语义索引模型、和支持向量机模型(SVM)等. VSM 模型将文档简化为以特征项的权重为分量的一个高维向量表示, 把文本信息过滤过程简化为空间向量的运算, 降低了问题复杂性, 可操作性好, 但是不能区分特征项出现在不同位置对表达文档主题性质能力的差异, 不能充分反映文本全貌, 而且特征项权重难确定. 贝叶斯决策模型能够解决预处理算法对多语种兼容性问题, 而且算法逻辑简单、易于实现、比较稳定, 较适用于垃圾邮件过滤领域, 但它没有考虑特征项在文档中出现的频率, 而且特征项是在独立性假设的基础上建立的, 所以错判率较高. 神经网络模型具有很强的自学习功能和自适应能力, 能够实现自我更新和完善, 但算法复杂、不支持部分匹配、执行速度慢, 不符合实时过滤的要求. 潜在语义索引模型能够保持特征项与文档之间的语义关系, 且能去除语言多义性等问题的影响, 但因算法复杂, 采用潜在语义的语义结构和大量新词的加入使过滤性能下降, 所以实际应用不多[4]. SVM 主要优势体现在解决高纬度、小样本以及线性不可分问题上, 目前已有学者将 SVM 应用到文本分类领域, 并取得了一定的进展, 但识别准确率与不良文本判定的召回率往往无法兼顾. 本文针对 SVM 判定能力与不良文本召回率无法兼顾的缺点, 提出了一种基于组合核函数的 SVM 方法, 通过多项式核函数与线性核函数的组合核函数, 来平衡单独的特征词汇以及多个特征词汇共现信息对文档不良性质判定的贡献, 从而使得更少的健康文本被错判为不良文本. 在提高语义判定能力的同时, 也兼顾不良文本判定的召回率.

2　SVM核函数原理

2.1 SVM 的基本思想

支持向量机(SVM)理论是 20 世纪 90 年代由 Vapnik 等人提出的一种新的机器学习方法[4], SVM 算法研究包含线性可分的问题和线性不可分的问题. 线性可分的问题是使用简单的线性分类器划分样本空间, 通过在特征空间构建具有最大间隔的最佳超平面得到两类文本之间的划分准则, 使期望风险的上限达到最小. 然而实际应用中, 大多数类别的文本并非相互隔离, 它们之间总有交界甚至彼此掺杂, 这就是线性不可分问题. 对于非线性问题, 可以通过非线性变换转化为某个高维空间中的线性问题, 在变换得到的维空间中寻求最优分类面. 引入核函数可以使SVM能够在尽量少的计算复杂度内得到高维映射后的特征向量内积, 从而可以使线性不可分问题转化为高维空间的线性可分问题进行求解.

2.2 核函数的基本原理

大量文本之间存在交界甚至掺杂使其无法用线性表述出来, 因此可以将原始数据向高维空间映射来增加线性函数的表达能力. 一般将可以接收低维空间向量并计算出映射后高维空间内积值的函数称为核函数, 即 $K(x,x')$.

假设 $(x_i,y_i),i=1,2\cdots K\cdots n,x\in R^d$, $y\notin\{1,-1\}$ 为样本集, 其中 y 是类别标号, 在一个 d 维空间中, SVM 处理非线性问题时, 首先通过一个非线性映射 $f:R^d\to H,x\to f(x)$, 将原空间 R^d 中的数据 x 映射到一个高维的特征空间 H 中, 通过引入核函数 $k(x_i,x_j)$[5], 使 SVM 可以在这个高维的核空间中使用线性函数将样本进行线性分类.

常见的核函数有四种: 线性内积核函数、多项式核函数、径向基核函数和二层神经网络核函数. 线性内积核函数为最简单的SVM核, 而多项式核函数由于表达简单, 参数较明确, 得到了广泛的应用. 本文将多项式核函数的思想应用于不良文本信息的识别, 并构造了新的核从而实现对有害文本特征组合语义信息的建模.

2.3 多项式核特征分析

多项式核函数的表现形式见公式(1).

$$K(x\cdot x_i)=\left[(x\cdot x_i)+C\right]^q,q>0 \tag{1}$$

上式表示的核是“q 阶多项式核”, 参数 q 被称为特征调节参数. 式中, $C\geq 0$ 是一个常数, 当 $C=0$ 时该核称为齐次多项式核, 当 $C>0$ 时该核称为非齐次多项式核. 在实际的应用中通常令 $C=0$.

多项式核函数可以实现将低维的特征空间映射到高维空间中, 从而实现线性不可分问题的求解[5]. 以二维样本空间为例, 从核函数的数学模型本身上来分析多项式核函数, 下式显示了该核在实现地维向高维特征空间映射的过程: 假设 $x\in R^2$, $X=[x_1,x_2]$ 当 $q=2$

时得到以下展开形式:

$$\left[(x_i \cdot x_j)+1\right]^2 = (x_{i1}x_{j1} + x_{i2}x_{j2} + 1)^2$$
$$= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 1 + 2x_{i1}x_{j1}x_{i2}x_{j2} + 2x_{i1}x_{j1} + 2x_{i2}x_{j2}$$
$$= (1, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}, \sqrt{2}x_{i1}x_{i2}, x_{i1}^2 x_{i2}^2) \cdot (1, \sqrt{2}x_{j1}, \sqrt{2}x_{j2}, \sqrt{2}x_{j1}x_{j2}, x_{j1}^2 x_{j2}^2)$$

以上所示情况对应的 SVM 的变换为公式(2):

$$\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)^T \quad (2)$$

此映射的主要作用是: 将二维空间上的曲线映射到 6 维空间上.

根据以上分析, 向量 $x \in R^2$ 被映射为 $\phi(x) \in H$, 则对于 $y = (y_1, y_2) \in R^2$, 上述 $q = 2$ 时的展开形式所示的相似性度量不仅考虑到了同一个特征词在两个文本中是否同时出现, 这对应用该核函数的SVM进行文本分类, 衡量两篇文档之间的差异是非常有意义的. 尤其当词汇的共现信息对文档的类别判定影响较大时, 使用该核函数应能取得更好的判定准确率.

3 不良词汇表的构建

在不良文本信息识别的应用中, 应该先构建不良样本词汇表, 主要应考虑以下三个因素.

①不良文本信息的特性. 因为绝大多数不良词汇是固定的, 尤其是色情、反动等类别的词汇, 因此, 词汇的收集工作一般独立于样本的语料来进行[5], 这些不良词汇可直接定性. 同时, 不良文本信息具有隐蔽性、多变性和多样性的特点, 导致有些词汇单独出现时, 不能直接判别为不良文档, 可将其划分为边界词汇种类.

②词表规模. 词表规模直接影响识别效果, 如果词表的规模过大, 虽然先验知识充分, 可以将含有不良词汇的样本都尽可能的识别出来, 有利于提高召回率, 但是大规模的词表更容易引入部分非不良词汇, 这样会导致准确率下降; 反之, 词表规模过小引起的问题是不良词汇表达不全面, 导致召回率降低.

③词条长度. 词条越长, 识别过程中完全匹配的几率就越小, 忽略了词条相互包含的关系, 因此会影响召回率; 词条过短, 只要包含该词条则被确定为有害是比较片面的, 因此会影响到识别的准确率.

在实际的识别过程中, 由于本文要兼顾准确率和召回率, 因此, 在构建特征词表时, 可构建大规模、稍短词条的词表. 词表中包含两种词汇, 一种为独立的不良词汇, 一种为边界词汇. 某些独立的不良词汇可以直接表征文档不良性质. 而对于边界词汇, 该词汇与其它某个或者某些词汇共同出现时, 才可以基本确定文档的倾向性, 这样也利于提高识别的准确率.

4 基于组合核函数的SVM核函数的构建

核函数的选择对于SVM是至关重要的. 已有不少理论及方法是关于单个特征空间的单核构造、改进或参数优化. 但不同的核函数具有不同的特性, 使其在不同的问题中性能表现差别很大, 特别是样本规模大、样本特征含有异构信息、数据在高维特征空间分布不平坦时, 采用单核并不合理. 而将两个或多个核函数组合可以兼顾各单核的优点[6]. 由于不同的核函数有不同的学习能力和推广能力, 组合核函数构成的SVM 兼有良好的学习能力和良好的推广能力. 而且通过调节组合核函数的核参数能调节组合核函数的性能, 可将最优核参数融入到核参数中.

在利用SVM核函数对特征词汇建模时, 本文采用组合核函数分别对独立的不良词汇以及多个不良词汇的组合信息进行建模, 达到不良信息识别的最终目的, 尽量减少边界样本的错判情况.

4.1 核函数构造规则

核函数可以有不同的组合方式, 但仍然需满足 Mercer 条件, 即假设 K_1 及 K_2 都是 $X \times X$ 上的核, $X \subseteq R^n$, 则对于 $\forall a \geq 0$, $K_1 + K_2$, $K_1 K_2$, aK_1 aK_1 都是核, 可以证明, 这些核都满足 Mercer 定理. 鉴于此, 可以将先验知识尽量融入到核函数的构造中, 组合后的核函数具有各种单核的优势, 能够得到性能更好的SVM[6].

4.2 一种新的组合核函数

对于多项式核函数, 若参数 C 为 0, 称为齐次多项式核, 若参数 C 不为 0, 称为非齐次多项式核, 具体描述如下.

①当 $C > 0$ 时

$$\phi_{ploy(q)}(x) = \left\{ \phi_I(x) = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \middle| I = (i_1, i_2, \cdots, i_n) \in N^n, 0 \leq \sum_{j=0}^{n} i_j \leq q \right\}$$

②当 $C = 0$ 时

$$\phi_{ploy(q)}(x) = \left\{ \phi_I(x) = x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} \middle| I = (i_1, i_2, \cdots, i_n) \in N^n, \sum_{j=0}^{n} i_j \leq q \right\}$$

上面两种多项式核都能刻画特征空间的组合信息, 区别在于是否包含阶数为 0 到 $q-1$ 阶的单项式, 从模型分析可以得到, 多项式核在对文档进行相似度度量时不仅考虑某一个特征词是否在两个文本中同时出

现，同时也考虑两个不同的特征词同时出现的影响，而且多项式核还实现了不良词语的组合信息建模，这样在新的特征空间中就能得到 q 个原始特征的组合特征. 基于线性核与齐次多项式核构造新的核函数，其定义公式(3):

$$K_{PL(q)}(x\cdot x_i)=\lambda(\sum_{d=2}^{q}K_{ploy(d)}(x\cdot x_i))+(1-\lambda)K_{linear}(x\cdot x_i) \quad (3)$$

其中，$K_{ploy(d)}(x\cdot x_i)$ 表示 d 阶齐次多项式核，$K_{linear}(x\cdot x_i)$ 为线性核，这种组合的核函数更加适合于不良信息过滤. 首先，组合的 $K_{PL(q)}$ 核函数是由两部分加权的和组成的:

① 第一部分: $\lambda(\sum_{d=2}^{q}K_{ploy(d)}(x\cdot x_i))$

q 阶以下的各阶数的特征组合，我们用来对不良词汇的组合行为进行建模;

② 第二部分: $(1-\lambda)K_{linear}(x\cdot x_i)$

对独立的不良词汇进行建模，表示不能忽略某些具有代表性的不良词汇的影响.

组合的核函数通过参数 λ 的调整，能够控制以上两部分在组合核函数中的作用程度，通过参数 q 能够控制多项式核的阶数，能够将不良词汇的组合规模限制在一定范围内.

由于不良词汇分为独立的不良词汇和边界词汇. 组合核函数由线性核和多项式核组合而成. 线性核插值能力较强，比较善于提取样本的局部性质，不良文本能够被准确的分类，因此比较适用于独立的不良词汇的判别. 多项式核函数插值能力弱，但推广能力强，善于提取样本的全局特性，因此比较适用于边界词汇的判别. 这些词汇与其他的某个或者某些词汇共同出现时，才可以基本确定文档的倾向性，需要考虑文档的全局特性. 组合核函数在线性核函数的作用下有较好的学习能力，单独识别误差较小，且在多项式核函数的作用下有很强的推广能力，全局识别误差小. 因此这个组合核函数具有较强的学习能力与推广能力. 应用这个组合核函数进行不良文本识别，不仅使独立的不良词汇的识别误差小，识别性能强，而且可以实现不良词汇组合进行判别，以确定文档的倾向性，兼顾了召回率和识别性能.

5　实验分析

为了验证本文算法的有效性，本文采用对比测试的方法进行实验，通过使用传统的SVM核与本文提出的改进的组合核函数分别进行了测试.

由于目前国内还没有标准、开放的不良文本语料集，因此，这里从 Internet 上收集与色情、反动、法论功等相关的不良文本 200 篇，人工标记为不良文本，其中一部分为通过多个边界词汇共现判定的不良文本. 从中文自然语言处理开放平台处下载了复旦大学的文本分类语料库，从中选取了包含经济、政治、军事三方面内容共 3800 篇正常文本，其中一部分文本包含容易错判的边界样本，它们包含了一些不良词汇，但是本身的语义倾向性是健康的. 将所有文本平均分成 4 组，每组包含 950 篇正常文本和 50 篇有害文本. 在仿真实验中对 4 组样本进行交叉测试，即第 1 组样本作为训练样本，随机抽取第 2 组样本中的 30%作为测试样本，其他各组样本测试方法类似，最后取多轮实验的结果的平均值.

实验采用中科院计算所汉语词法分析系统 ICTCLAS 进行分词，使用停用词过滤和词根萃取的方法去除冗余特征. 用两步特征选择方法得到特征集，对训练文本集中的每篇文本的特征项用 TFIDF 函数计算特征项权值，并以向量模式表示语料库的样本点[7].

评价标准采用 Recall(召回率)、precision(准确率)以及 F-value(标准测度)，其中 F-value 是综合了 Recall、Precision 的评价指标，见公式(4)，针对本文的特定应用，令参数 $\beta=2$.

$$F-value=\frac{(1+\beta^2)\times Recall\times Precision}{\beta^2\times Precision+Recall} \quad (4)$$

在仿真实验评估了线性核、多项式核以及本文提出的组合核函数，参数 λ 、q 取不同的值进行实验，以衡量参数的值对组合核性能的影响，多项式核阶数最高为 4 阶，实验结果如表 1 所示.

表 1　不同 SVM 核函数识别效果

核函数	参数	Recall	precision	F-value
线性核	无	0.781	0.862	0.842
齐次多项式核	q=2	0.702	0.915	0.790
	q=3	0.651	0.931	0.756
	q=4	0.689	0.927	0.780
组合核函数	q=3, $\lambda=0.4$	0.862	0.942	0.884
	q=3, $\lambda=0.7$	0.877	0.961	0.898
	q=3, $\lambda=0.9$	0.881	0.978	0.918
	q=4, $\lambda=0.9$	0.847	0.973	0.871

从线性核和多项式核的实验数据来看，准确率和

召回率是成反比的, 准确率越高, 召回率则越低. 结果显示两者识别准确率较高, 但召回率相对较低. 这是由 SVM 本身的特点所决定的, 因为 SVM 是以结构风险最小化原则为理论基础的一种算法, 目标就是保证识别准确率, 从而无形中影响召回率.

线性核的识别性能较差, 说明仅使用独立的不良词汇作为特征无法满足不良信息过滤的要求. 多项式核的识别准确率要明显高与线性核, 这是因为多项式核能够实现不良词汇的组合信息建模, 但召回率较低. 同时可以看出多项式核的性能参数 q 对 SVM 的性能整体影响不大, 这是因为多项式核是全局核函数. 而组合核函数在 Recall、precision 以及 F-value 三个性能评价指标上普遍优于线性核和齐次多项式核. 这是由于组合核函数兼顾了线性核较强的学习能力与多项式核较强的推广能力, 对独立的不良词汇与词汇组合都达到较高的识别率, 兼顾了召回率和识别准确率. 特别是当 $q=3$, $\lambda=0.9$ 时, Recall、precision 以及 F-value 三个性能评价指标均达到最佳, 而且 q 对组合核函数的性能影响不大, 而 λ 的取值对识别结果影响较大, 当 λ 取值为 0.9 时, 综合性能最佳, 这是由于在组合核函数中突出了对特征共现信息建模的部分, 对 SVM 分类性能影响较大.

6 结语

SVM 中最关键的问题就是找到适合的核函数, 不同的核函数将形成不同的算法. 本文基于现有的核函数, 依据 Mercer 规则构造了一种新的 SVM 组合核函数, 结合线性核与多项式核, 分别对独立的不良词汇以及组合语义信息进行建模, 新的组合核函数在不良文本信息过滤的实验中取得较好的性能, 但这种组合的核函数也存在一定的问题, 如组合的核函数可能引入更多的参数, 而参数的选择会影响 SVM 性能, 同时使模型选择的困难增大, 这个问题有待下一步研究.

参考文献

1 曾铭,俞俊生,刘绍华.一种用于社交网站的云安全敏感信息过滤模型.华中科技大学学报(自然科学版),2012,(S1):211–214.

2 王子强,张文阁,王洪艳.基于内容的网络异常信息过滤.硅谷,2012,9(18):9–10.

3 Theodoridis S, Koutroumbas K.李晶皎等译.模式识别.3rd ed.北京:电子工业出版社,2006.

4 丛健.不良信息过滤技术研究[硕士学位论文].北京:北京邮电大学,2012.

5 杜娟.基于内容的网络信息过滤模型的应用研究[硕士学位论文].大庆:大庆石油学院,2009.

6 瞿娜娜.基于组合核函数支持向量机研究及应用[硕士学位论文].广州:华南理工大学,2011.

7 苏红刚.基于 SVM 的中文文本分类系统实现[硕士学位论文].长春:吉林大学,2012.