组合核函数 SVM 在特定领域文本分类中的应用①

吕洪艳, 刘 芳

(东北石油大学 计算机与信息技术学院, 大庆 163318)

摘 要: 面向特定领域文本分类的实际应用, 存在大量样本相互掺杂的现象, 使其无法线性表述, 在SVM中引入核函数可以有效地解决非线性分类的问题, 而选择不同的核函数可以构造不同的 SVM, 其识别性能也不同, 因此, 选择合适的核函数及其参数优化成为 SVM 的关键. 本文基于单核核函数的性质, 对多项式核函数与径向基核函数进行线性加权, 构建具有良好的泛化能力与良好的学习能力的组合核函数. 仿真实验结果表明, 在选择正确参数的情况下, 组合核函数 SVM 的宏平均准确率、宏平均召回率及宏平均综合分类率都明显优于线性核、多项式核与径向基核, 而且能够兼顾准确率与召回率.

关键词: SVM; 组合核函数; 文本分类; 多分类

Application of Text Classification for Specific Domains Based on Combination Kernel Function SVM

LV Hong-Yan, LIU Fang

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

Abstract: In practical application of text classification for specific domains, most of the text always dopes with each other and is unable to be expressed in linear. The application of kernel function in SVM can solve the problem of nonlinear classification effectively. Different SVM can be constructed by different kernel function, and the recognition performance is also different. So the key problems of SVM are choosing the appropriate kernel function and optimizing its parameters. This paper constructs a new combination kernel function combined with homogeneous polynomial kernel and radial basis kernel function by linear weighted method based on the character of the kernel function. The combination kernel function has good generalization ability and good earning ability at the same time. The simulation experiment result shows that the precision rate, recall rate and comprehensive classification rate of macro average of combination kernel function are obviously better than linear kernel, polynomial kernel and radial basis kernel in choosing the right parameters, and the precision rate and the recall rate are ideal.

Key words: SVM; combination kernel function; text classification; multi-classification

1 引言

随着互联网的迅猛发展, 信息增长速度越来越快, 使得信息量也非常巨大. 在一般情况下, 特定用户只需求某一领域内的信息, 而在海量信息中寻找所需信息需要耗费大量的时间与精力, 因此, 为了使特定领域的人们能够高效地获得所需信息, 就需要对信息进行自动分类. 由于互联网上的大部分信息都以文本形式存在, 因此, 文本信息自动分类就显得尤为重要. 针对特定领域的用户需求, 基于内容的文本自动识别与分类方法是组织、管理文本信息的有效手段, 可以有效地提高用户寻找所需信息的效率.

国外关于文本自动分类研究的较早, 20 世纪 50 年代末, H.P.Luhn 就对文本自动分类进行了开创性的研究. 此后, At&T 实验室、美国 Just Research 公司、德国 Dortmund 大学计算机系等多个单位、企业进行深入研究, 并研制出一批自动分类系统[1]. 国内对其研究起步较晚, 候汉清教授在 1981 年首先对自动文本分类进行了探讨. 此后, 南京大学、中国科学院、复旦大学、清华大学等单位开始从事该领域的研究, 并研制出了一批自动分类系统[1]. 目前, 国内外对于文本自动分

① 收稿时间:2015-08-30;收到修改稿时间:2015-10-30

类的研究重点主要集中在文本分类的算法与模型上. 主要研究的模型有向量空间模型(VSM)、贝叶斯决策模型、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, 降低了问题复杂性, 但是不能充分反映文本全貌, 而且特征项权重难确定[2]. 贝叶斯决策模型算法逻辑简单且比较稳定, 但它的特征项是在独立性假设的基础上建立的, 所以错判率较高. 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, 但由于算法复杂和大量新词的加入使识别性能不佳, 所以实际应用不多[2]. 支持向量机(SVM)是由 Vapnik 等人在 1996 年提出的基于结构风险最小化原理的一种机器学习方法, 得到的是全局最优解, 这使得它有着其它统计学习技术难以比拟的优越性. 其主要优势体现在解决高纬度、小样本以及非线性分类问题上, 缺点是选择合适的核函数及其参数比较困难, 而且利用单核函数无法兼顾识别准确率与召回率.

目前, 文本分类在多种领域都起着至关重要的作用. 但依然存在的主要问题有小样本问题、非线性文本分类问题, 识别准确率与召回率无法兼顾、样本分布不均衡等问题[3]. 这些问题影响了一些文本自动分类系统的应用范围甚至降低其识别性能, 成为目前文本识别领域研究的热点, 也是亟待解决的问题. 针对前三个问题, 这里选择 SVM 分类器, 解决小样本及非线性分类问题, 在此基础上将多项式核函数与径向基核函数进行组合, 并运用多重网格搜索法进行参数优化, 以期解决准确率与召回率无法兼顾的问题, 提高文本主题识别性能.

2 基于SVM的文本信息过滤的流程

基于 SVM 的文本分类过程分为训练阶段和分类阶段. 由于多数 Web 网页中包含广告、注释等与内容无关的信息, 因此, 为了方便后续处理, 需将文本进行预处理, 也就是将无关信息去除. 在训练阶段, 将分词结果存入特征库[4], 根据特征库中的特征项进行特征提取, 得到的是重要的特征项, 依据重要特征项通过一定的方法计算出特征权重, 并表示成向量形式, 在此基础上训练 SVM 分类器. 在分类阶段, 对待测样本进行预处理、分词处理后, 再进行特征提取, 计算出特征权值, 并表示成向量[4], 再运用 SVM 进行分类, 判定待测样本的类别, 具体流程见图 1.

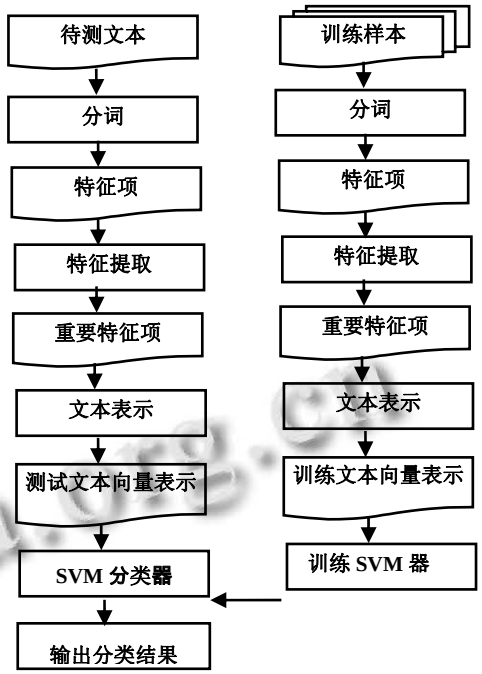

图 1 基于 SVM 的文本信息过滤过程

(1) 分词. 经过预处理后的训练样本集中的文本需进行分词处理, 文本经过分词后可以表示成词语的组合. 这些词语不可能都作为特征表示文档, 此时, 可以用特征项来代表文档的内容, 特征项的集合形成特征库. 这里采用由中科院计算所开发的汉语词法分析系统 ICTCLAS 进行分词.

(2) 特征提取. 从特征集中选取部分最重要的特征就是特征提取, 特征提取同时也能降低样本空间的维度. 这里采用 CHI 统计法进行特征提取.

(3) 文本表示. 文本表示是将文本抽象成某种数学模型的过程, 这里采用VSM模型中最为经典的特征权值函数(TFIDF)将文本表示城向量形式.

(4) SVM 分类器.

SVM 的基本思想是寻找一个最优分离超平面, 把两类样本正确分开, 使类内差别最小, 分类间隔最大[5]. 具体步骤如下:

①构建最优分类面 $y(x)=\mathrm{sgn}\{w\bullet\phi(x)+b\}$ 将训练样本向量集分开, 其中 x 是输入样本的向量形式, w 是超平面的法向量, b 为偏移量, $\phi(x)$ 表示特征映射.

②将最优分类面问题转化为求解以下问题:

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

$$s.t.\ y_i\left(w^t\bullet\phi\left(x_i\right)+b\right)\geq 1-\xi_i,\xi_i\geq 0,i=1,\ldots,N$$

其中 C 是对错分样本的惩罚因子，ξ_i 是松弛变量.

③应用Lagrange将以上问题转化为对偶问题:

$$\max Q(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jy_iy_j\varphi(x_i)\varphi(x_j) \quad (2)$$

$$s.t.\ \sum_{i=1}^{N}\alpha_iy_i=0,0\le\alpha_i\le C,i=1,2,\ldots,N$$

其中 α_i 为Lagrange乘子.

④引入满足 $K(x_i,x_j)=\phi(x_i)\bullet\phi(x_j)$ 的函数(核函数)解决非线性问题. 相应的二次规划问题为:

$$\max Q(a)=\sum_{i=1}^{N}a_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_ia_jy_iy_jK(x_i,x_j) \quad (3)$$

⑤求解以上问题，得到解 $a'=(a'_1,a'_2,\cdots,a'_n)^T$，选择 a' 中大于零的分量代入 $b'=y_i-\sum_{i=1}^{n}y_ia'_iK(x_i,x_j)$ 中得到阈值b.

⑥将测试文本集中文本 x_i 代入构造的分类函数 $f(x)=\mathrm{sgn}(\sum_{i=1}^{n}y_ia'_iK(x_i,x_y)+b)$ 中，来判定文本类别.

在文本识别的实际应用中，大量文本之间总是彼此掺杂，使其无法线性表述，核函数的引入可以使高维空间的内积运算转化为原空间一个内积核函数的计算[5]，在不增加算法复杂度的同时实现了非线性算法.

3　SVM 核函数

3.1 SVM 核函数的主要形式及其性质

选取不同的核函数可以构建不同 SVM，因此，核函数的选择至关重要. 常见的核函数有以下四种.

①线性内积核函数

$$K(x_i\cdot x_j)=(x_i\cdot x_j) \quad (4)$$

②多项式核

$$K(x_i\cdot x_j)=[(x_i\cdot x_j)+C]^q,q>0 \quad (5)$$

在公式(6)中，q 是多项式核函数的幂指数，$C\ge0$ 是一个常数，在实际应用中通常令 $C=1$. 不同类型的核函数表现出不同的特性，根据其特性不同，常分为局部性核函数和全局性核函数. 多项式核函数是典型的全局核函数，它作用范围较广，甚至对整个数据点都有影响，而且对于离测试点较远的数据仍然有较强的影响，泛化能力较强，但局部学习能力较弱[6].

③径向基核

$$K(x_i\cdot x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2) \quad (6)$$

上式中，σ 是径向基核函数的宽度系数. 径向基核函数是典型的局部核函数，由其表达式可知，当 $\sigma\to0$ 时，全部训练样本都能够正确分类，但它只对样本距离 σ 相当的小范围内的样本有效，当样本距离大于 σ 且逐渐增大时，它的核函数值会逐渐下降，而且下降的速度会越来越快，说明径向基核函数局部性较强，泛化能力较弱.

④两层神经网络(Sigmoid 核)

$$K(x_i\cdot x_j)=\tanh((v(x_i\cdot x_j)+\theta) \quad (7)$$

v 是一个标量，θ 是其位移参数. Sigmoid 核函数由于只有参数满足特定条件时，才是半正定的，所以在实际应用不多，这里也不考虑 Sigmoid 核函数.

3.2 组合核函数构建

由于不同的核函数具有不同的特性，使其在解决具体问题时表现差别很大. 目前，关于单核的构造、改进及其参数优化的研究较多，但采用单核SVM的识别效果并不理想. 所以，组合核函数是目前选取核函数的最佳方法. 由上可知，全局核函数泛化能力较强，比较适合提取样本的全局特性，而局部核函数学习能力较强，比较适合于独立词汇的判定，因此，可以将这两种核函数进行组合，充分发挥它们各自的优点，使组合核函数兼具良好泛化能力与良好学习能力，以提高文本识别性能.

核函数可以有不同的组合方式，但仍然需满足 Mercer 条件. 若 K_1 及 K_2 都是 $X\times X$ 上的核，$X\in R^n$，则可以证明 $\forall a,b\ge0$，，核 aK_1+bK_2 也满足 Mercer 定理. 这里将多项式核与径向基核进行线性组合，即:

$$K(x_i\cdot x_j)=\alpha[(x_i\cdot x_j)+1]^q+(1-\alpha)\exp(-\|x_i-x_j\|^2/2\sigma^2) \quad (8)$$

其中 $\alpha>0$，表示两种核函数的比例系数.

3.3 组合核函数 SVM 参数优化方法

SVM 核函数参数优化的主要方法有交叉验证法、网格搜索法及遗传算法、粒子群算法、蚁群算法等一些群智能算法. 交叉验证法优点是简单、易于实现，缺点是精度不高. 网格搜索法优点是模型简单且可以同时搜索多个参数，缺点是当参数较多时精度不高，但多重网格搜索法可以在一定程度上解决这个问题. 群智能算法复杂度相对要高得多，此外，遗传算法的优点是对目标函数要求不高，缺点是受初值影响较大. 粒子群的优点是收敛速度快，缺点是精度不稳定. 蚁群算法的优点是精度较高，缺点是收敛速度慢和鲁棒性差[7]. 综合考虑算法复杂度和参数精度，这里选择

多重网格搜索法. 网格搜索法的主要思路是首先选取合适的搜索范围, 然后确定搜索的步长, 以固定的步长沿着各个参数方向生成网格, 网格中的节点就是初始给定范围内的所有可能的参数组合. 多重网格搜索法从上一次网格寻优确定的最优点开始, 再次进行网格寻优, 减小搜索步长, 以此类推. 构造的组合核函数中共有惩罚因子 C、多项式核函数的幂指数 q、径向基核函数的宽度系数 σ、及比例系数 α 四个参数. 如要确定参数 C 与 q, 首先设定参数 C 的范围为 $C\in[C_1,C_2]$, 搜法步长为 C_s, 参数 q 的范围为 $q\in[q_1,q_2]$, 搜法步长为 q_s, 然后针对每对参数 $[C',q']$ 进行训练. 多重网格搜索法是完成一次网格搜索后得到一组最优的参数组合 $[C'',q'']$, 再对 $[C'',q'']$ 附近一定范围内实现更细致的网格搜索, 以提高参数优化精度.

3.4 SVM 多分类方法选择

针对 SVM 在特定领域文本分类中的特定应用, 需解决 SVM 的多分类问题, 常用的多分类方法有“一对多”方法、“一对一”方法和有向无环图(DAG). 由于“一对多”方法和“一对一”方法分别存在多个得票数最多和多个最大决策函数输出值的情况, 使得误分率较高, 所以这里选择 DAG 策略. 应用 DAG 方法将待测样本分类时, 从顶部节点开始, 根据判别结果沿着无环图从顶层向下一层游走, 直到在底层找到待测样本的所属类别[8].

4 实验分析

4.1 数据来源

本文采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供的文本分类语料库[9]. 该语料库共收集了 19637 篇文本, 测试语料有 9833 篇, 训练语料有 9804 篇, 均分 10 类. 这里抽取的文本类别及数量如表 1 所示.

表 1 实验语料类别及数量

	交通	医药	政治	军事	体育	环境	教育	经济
训练样本数	150	150	400	200	320	140	180	280
测试样本数	80	70	200	100	150	60	80	130

4.2 最优参数确定

对于组合核函数, 根据选定的多重网格搜索法, 第一次搜索 C、σ、α、q 的范围分别是[1, 600], [0, 10], [0, 1], [1, 20], 步长分别是 10, 0.5, 0.1, 1, 得到的最优参数为 $C=120$, $\sigma=0.5$、$\alpha=0.2$、$q=3$. 在此基础上进行再次优选, C、σ、α、q 的范围分别是[80, 160], [0, 2], [0, 0.4], [1, 6], 步长分别是 1, 0. 05, 0.01, 1, 得到的最优参数为 $C=128$, $\sigma=0.2$、$\alpha=0.15$、$q=3$. 便于比较实验结果, 需确定单核核函数参数如下: 线性核参数 $C=64$, 多项式核参数 $q=3$, $C=16$, 径向基核函数参数 $\sigma=0.25$, $C=8$.

4.3 性能评价指标

评价标准采用查全率(Recall 用 R 表示)、准确率(Precision 用 P 表示)以及综合分类率(F-value 用 F 表示), 并采用宏平均(Macro)值即宏平均查准率(MacP)、宏平均查全率(MacR)和宏平均综合分类率(MacF)来评价文本分类系统的整体分类性能. 假设用 N_1 表示被正确分类的文本数, N_2 表示某类文本的实际数, N_3 表示归入某类文本(属于该类与不属于该类的文本)的总数, 则可得准确率、查全率和综合分类率分别为 $P=N_1/N_2$、$R=N_1/N_3$ 和 $F=P\times R\times 2/(P+R)$, 相应的宏平均准确率、查全率和综合分类率分别为 $MacP=\frac{1}{n}\sum_{i=1}^{n}P_i$、$MacR=\frac{1}{n}\sum_{i=1}^{n}R_i$ 和 $MacF=\frac{MacP\times MacR\times 2}{MacP+MacR}$.

4.4 实验结果与对比分析

应用组合核函数 SVM 及其最优参数组合对测试语料进行分类, 得到的实验结果见表 1 如表 2.

表 2 组合核函数 SVM 分类结果

		被分到的类								
		交通	医药	政治	军事	体育	环境	教育	经济	总数
样本原类别	交通	77			1		1		1	80
	医药		68	1			1			70
	政治			180	15		1	2	2	200
	军事	1		10	88			1		100
	体育	1				149				150
	环境			2			56	1	1	60
	教育		1			1	1	76	1	80
	经济	1		2			1	1	125	130
总数		80	69	195	104	150	61	81	130	

表 3 组合核函数 SVM 分类性能

	交通	医药	政治	军事	体育	环境	教育	经济	宏平均
P	0.963	0.971	0.900	0.880	0.993	0.933	0.950	0.962	0.944
R	0.963	0.986	0.923	0.846	0.993	0.918	0.938	0.962	0.941
F	0.963	0.978	0.911	0.863	0.993	0.926	0.944	0.962	0.943

由表 2 和表 3 可知, 在各类的准确率、查全率与综合分类率中, 体育类的最高, 值都是 99.3%, 军事类

的最低, 值分别为 88%、84.6%和 86.3%, 最高值与最小值之间分别相差 11.3%、14.7%和 13%, 波动较大. 这是由于军事类有 10 篇被分到政治类, 而政治类有 15 篇被分到了军事类, 导致军事类和政治类的准确率与查全率都较低, 因为从语义概念分析, 政治与军事两类本身就有较高的相似度. 准确率、查全率与综合分类率的宏平均值分别为 94.4%、94.1%和 94.3%, 最大值与最小值的平均值分别为 93.65%、91.95%和 92.8%, 之间相差分别为 0.65%、2.15%和 1.5%, 相差较小, 说明除了语义概念较相似的军事类和政治类外, 系统在准确率、查全率和综合分类率方面的分布不存在过分集中在某特定类的现象.

分别以线性核、多项式核、径向基核以及本文提出的组合核函数作为分类器, 以比较它们的识别性能, 并针对组合核函数选取不同的参数组合, 以比较参数对组合核函数 SVM 分类性能的影响. 实验结果见表 4.

表 4　不同 SVM 核函数识别效果

核函数	参数	MacR	MacP	MacF
线性核	C=64	0.768	0.712	0.739
多项式核	q=3, C=16	0.684	0.854	0.760
径向基核	σ =0. 2, C=8	0.737	0.812	0.773
组合核函数	σ =0.2, α =0.15, q=3, C=128	0.941	0.944	0.942
	σ =0.25, α =0.15, q=3, C=128	0.795	0.905	0.846
	σ =0.2, α =0.3, q=3, C=128	0.782	0.851	0.815
	σ =0.2, α =0.15, q=4, C=128	0.892	0.934	0.913
	σ =0.2, α =0.15, q=3, C=64	0.884	0.792	0.835

由表 4 可知, 以线性核、多项式核和径向基核构建的单核 SVM 中, 在选择最优参数情况下, 宏平均准确率分别为 71.2%, 85.4%和 81.2%, 线性核的识别性能最差, 多项式核识别平均准确率最高, 但总体来说, 准确率都不高. 宏平均召回率分别为 76.8%, 68.4%和 73.7%, 说明准确率和召回率是成反比的, 准确率越高, 召回率则越低. 结果显示三者识别准确率相对较高, 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.

对于组合核函数, 在最优参数的基础上, 分别改变各参数的值, 得到的结果也都不理想, 甚至有低于两个单核的准确率的情况. 这些都说明如果参数选择不准确, 组合核函数 SVM 也无法得出理想的结果. 若选择的正确的参数, 组合核函数 SVM 的宏平均准确率为 94.4%, 宏平均查全率为 94.1%, 宏平均综合分类率为 94.3%, 三个性能评价指标均达到最佳, 远高于其他三个单核的性能指标, 这是由于组合核函数兼顾了径向基核函数较强的学习能力与多项式核函数较强的推广能力, 能够兼顾文本分类的准确率和召回率.

5　结语

核函数的引入在不增加算法复杂性的情况下, 解决了实际应用中广泛存在的非线性分类问题. 由于单核函数无法兼顾识别准确率和召回率, 本文根据单核的特点构造了组合核函数. 在仿真实验中评估了线性核、多项式核、径向基核以及组合核函数的分类性能, 实验结果表明, 组合核函数 SVM 的识别性能明显优于其它三个单核函数. 由于语料库的质量直接影响分类性能, 但目前并不存在一个标准的中文分类语料库, 因此, 本文的结论还需要更多的实验来验证, 且如何减少分类性能对语料库的依赖性也是要研究的问题.

参考文献

1 褚金正.面向特定领域的文本识别和分类[硕士学位论文].长沙:湖南大学,2005.

2 吕洪艳,杜娟.基于 SVM 的不良文本信息识别.计算机系统应用, 2015,24(6):183−187.

3 胡明涵.面向领域的文本分类与挖掘关键技术研究[博士学位论文].沈阳:东北大学, 2009.

4 庄新妍.基于 SVM 的中文文本分类系统的研究与实现[硕士学位论文].长春:吉林大学, 2007.

5 高会生,郭爱玲.组合核函数 SVM 在网络安全风险评估中的应用.计算机工程与应用,2009,45(11):123−124.

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 杨海.SVM 核参数优化研究与应用[硕士学位论文].杭州:浙江大学,2014.

8 叶志刚.SVM 在文本分类中的应用[硕士学位论文].哈尔滨:哈尔滨工程大学,2006.

9 李荣陆.复旦大学提供的文本分类语料库.http://www.datatang.com/data/43543.[2012-8].