SVM-KNN组合改进算法在专利文本分类中的应用

李程雄[1] 丁月华[2] 文贵华[2]
[1](广东粤华发电有限责任公司信息分部,广州 510731)
[2](华南理工大学计算机应用工程研究所,广州 510640)
E-mail:Lcx1218@21cn.com

摘 要 提出了基于支持向量机的专利文本分类器的总体设计方案和实现方法;提出并分析了该分类器的改进算法SVM-KNN组合改进算法。文章对两种算法进行了大量的实验并对实验结果进行比较分析,在此基础上得出了三个结论。

关键词 支持向量机 KNN 专利分类 最优分类面

文章编号 1002-8331-(2006)20-0193-03 **文献标识码** A **中图分类号** TP181

Application of SVM-KNN Combination Improvement Algorithm on Patent Text Classification

Li Chengxiong[1] Ding Yuehua[2] Wen Guihua[2]
[1](Information Division, Guangdong Yuehua Power Company LTD., Guangzhou 510731)
[2](Research Institute of Computer Application, South China University of Technology, Guangzhou 510640)

Abstract: It narrates the overall design plan and implementation method of patent text classification machine resulting from support vector machine; proposes and analyzes its improvement algorithm SVM-KNN combination improvement algorithm; and a great deal of tests on classification machine are carried out to two algorithms and the testing results are compared and analyzed, draws three conclusions in this foundation.

Keywords: support vector machine, KNN, patent classification, optimal hyperplane

在当今全球化经济的时代,专利技术已成为国家或地区竞争力的核心,专利知识产权越来越受到企业的重视。因此,近几年的专利申请量迅速增长,但是目前专利分类仍是采用传统的手工分类,这种分类的方法效率低下,存在许多弊端,如周期长、费用高、效率低,分类结果一致性不高等问题。专利申请量的激增一方面增加了对快速、自动文本分类的迫切需求,另一方面又为基于数据挖掘技术的文本分类方法准备了充分的资源。因此,计算机辅助专利分类成为大势之所趋[1]。

当前对支持向量机的研究是一个热点,支持向量机是基于统计学习理论的机器学习方法,有一套坚实的理论基础。遗憾的是,虽然支持向量机在理论上有很突出的优势,但与其理论研究相比,应用研究尚相对比较滞后,目前只有比较有限的实验研究报道,且多属仿真和对比实验。

研究目前利用支持向量机实现文本分类的现状可以发现,虽然存在很多这样的应用系统,但分类对象都是新闻资料或网页资料,而对于应用于中国专利的分类则还没有,所以这也是本文的一个创新点[2]。

但目前在SVM的应用中还存在一些问题,如对不同的应用问题核函数参数的选择较难,对较复杂问题其分类精度不是很高以及对大规模分类问题训练时间长等。已有的解决方法包括建立分类性能的评价函数,然后对SVM中的核函数的参数进行优化,或者使用直推方法对给定待样本设计最优的SVM;所有这些方法的设计和计算都非常复杂,实现的代价都很高。因此系统采用了SVM-KNN组合算法对分类器进行改进,并取得了一定的效果。

1 SVM-KNN(KSVM算法)组合改进算法介绍

有关支持向量机的基本知识和原理可以参考文献[3~6]。

近邻法(简称NN)是模式识别非参数法中最重要的方法之一,NN的一个很大特点是将各类中全部样本点都作为"代表点"。1NN是将所有训练样本都作为代表点,因此在分类时需要计算待识别样本x到所有训练样本的距离,分类结果就是与x最近的训练样本所属于的类别。KNN是1NN的推广,即分类时选出x的k个最近邻,看这k个近邻中的多数属于哪一类,就把x分到哪一类。

我们对SVM分类时错分样本的分布进行分析发现,SVM分类器和其它的分类器一样,其出错样本点都在分界面附近,这提示我们必须尽量利用分界面附近所提供的信息以提高分类性能。由SVM理论知道,分界面附近的样本基本上都是支持向量,同时SVM可以看成每类只有一个代表点的最近邻(Nearst Neighbour, NN)分类器。所以结合SVM和NN,对样本在空间的不同分布使用不同的分类法。具体地,当样本和SVM最优超平面的距离大于给定的阈值,即样本离分界面较远,则用SVM分类,反之用KNN对测试样本分类。在使用KNN时以

基金项目:国家自然科学基金资助项目(编号:60003019)

作者简介:李程雄(1979-),男,硕士研究生,研究方向为智能技术与数据挖掘。丁月华(1955-),男,教授,博士生导师,华南理工大学计算机应用研究所所长,研究方向为计算机应用技术。文贵华(1968-),男,博士,硕士生导师,研究方向为人工智能、计算机应用。

每类的所有的支持向量作为代表点组，这样增加的运算量很少。实验证明了使用支持向量机结合最近邻的分类器分类比单独使用支持向量机分类具有更高的分类准确率，同时可以较好地解决应用支持向量机分类时核函数参数的选择问题。

将 SVM 和 KNN 分类器结合的考虑是将 SVM 看成每类只有一个代表点的 1NN 分类器。由于 SVM 对每类支持向量只取一个代表点，有时该代表点不能很好的代表该类，这时将其与 KNN 相结合是因为 KNN 是将每类所有支持向量作为代表点，从而使分类器具有更高的分类准确率。具体地，对于待识别样本 x，计算 x 与两类支持向量代表点 x+和 x- 的距离差，如果距离差大于给定的阈值，即 x 离分界面较远，如图 1 中区域 I 和 II，用 SVM 分类一般都可以分对。当距离差小于给定的阈值，即 x 离分界面较近，即落入区域 III 时，如分类用 SVM，只计算 x 与两类所取的一个代表点的距离比较容易分错，这时采用 KNN 对测试样本分类，将每个支持向量作为代表点，计算待识别样本和每个支持向量的距离对其得出判断。在对数据的封闭测试（用训练样本作为测试集）中，SVM- KNN 分类器的输出接近于 100%。这是由于支持向量大多位于分类超平面附近，即属于上图中的区域 III，此时代入 1NN 对其分类，对于每个支持向量，都能找到支持向量自己作为最近邻，其结果总是正确的[7,8]。

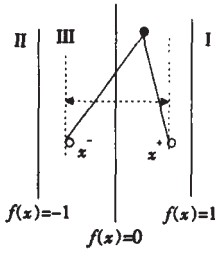

图 1 KNN 分类说明

2 基于 SVM 的专利文本分类系统的设计与实现

2.1 专利分类设计的背景

国际专利分类系统分成部、大类、小类、大组和小组等 5 级结构，涉及与发明专利有关的全部技术领域。它分八大部，用 A, B, C, D, E, F, G, H 等 8 个大写拉丁字母分别表示；部以下有大类（class），用阿拉伯数字表示，如 A24；大类以下是小类（subclass），以大写拉丁字母表示，如 A24D；小类以下则有大组（group）和小组（sub- group），用阿拉伯数字表示，在大组和小组之间用斜线“/”隔开，如大组 A24D3/00 和小组 A24D3/12。

目前专利分类是采用传统的手工分类，工作人员根据专利信息内容，对照国际分类表，手工检索出其所属的部、大类、小类、大组和小组。这种分类的方法效率比较低下，而且很受工作人员的主观因素影响，准确率也相对比较低，因此，计算机辅助专利分类势在必行。

2.2 SVM 专利分类系统的设计

专利自动分类模块主要由训练和分类两部分组成，分别对应文本分类的训练和分类这两个阶段。训练部分是对训练样本进行文本分词、词语处理及特征抽取、生成训练集特征向量集、参数训练，生成分类模型；分类部分首先对要识别的样本进行分词、计算特征词权重、生成测试集特征向量集，然后根据训练所得的分类知识库，通过 SVM 识别算法对样本进行自动分类。其框架体系如图 2 所示，其中文本内容主要由专利信息中的专利名称和摘要两部分组成。

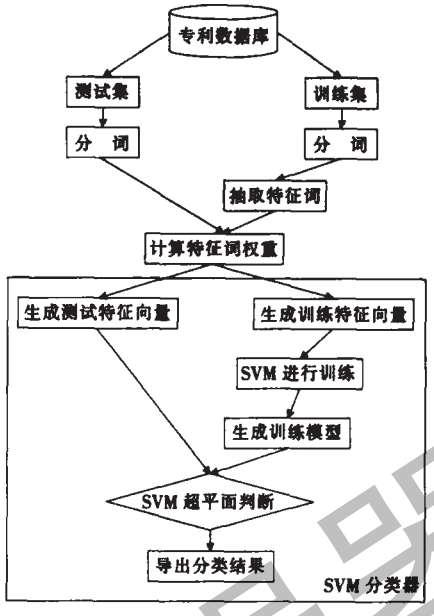

图 2 专利自动分类子系统的框架体系

该专利文本分类系统主要分成两部分，一部分是分类前的数据准备部分；另一部分是分类部分，总体设计如图 3 所示。

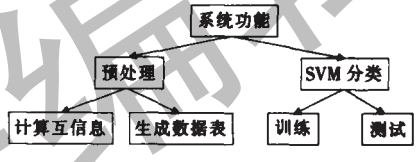

图 3 专利自动分类子系统的系统功能

2.3 分类算法的实现

上面系统的框架说明了分两类的算法：

（1）选择特征词：先在数据库中选出样本集进行分词，然后计算各词语的互信息，互信息的计算公式是：

$$MI(w,c)=\log\frac{X\times N}{(X+Z)\times(X+Y)} \quad (1)$$

其中 c：类别，w：词语，X：w 出现在类中的次数，Y：w 出现在非类中的次数，Z：类中不出现的次数，N：总的文档数。最后在两个类中选出等数量的并且不重复的互信息值最高的 m 个特征词；

（2）构成训练集特征向量表：计算特征词权重：$w_c=\lg(2T_c+A_c+1)$，w_c 是该词的权重，T_c 是该词在专利名称中出现的次数，A_c 是该词在摘要中出现的次数；以样本集中的文档为行，文档中各特征词为列构成表示训练集各文档各特征词权重的特征向量表；

（3）把特征向量表导入 SVM 两类分类器进行训练，形成这两类样本的训练模型；

（4）按照同样的方法构造测试集的特征向量表，导入分类器利用训练模型对其进行分类，最后导出测试结果，结果有三种情况，第一种情况是无法识别的，另两种情况就是经过 SVM 分类器把文档分成了两个类别。

3 SVM- KNN 改进算法

首先利用任何一种 SVM 算法，求出相应的支持向量和它的系数以及常数 b，设 T 为测试集，Tsv 为支持向量集，k 为

KNN 的个数。

步骤 1 如果 $T\neq\Phi$, 取 $x\in T$, 如果 $T=\Phi$, 停止;

步骤 2 计算公式: $g(x)=\sum_i y_i\alpha_i k(x_i,x)+b$ (2)

步骤 3 如果 $|g(x)|>\varepsilon$, 直接计算 $f(x)=\mathrm{sgn}(g(x))$ 作为输出; 如果 $|g(x)|<\varepsilon$ 代入 KNN 算法分类, 传递参数 x、Tsv、k, 返回结果为输出。

步骤 4 T=T-{x}, 返回步骤 1。

上述算法步骤 3 中使用的 KNN 分类算法和通常的 KNN 分类算法流程相同, 将支持向量集 Tsv 作为分类算法的代表点集合即可。所不同之处在于计算测试样本和每个支持向量的距离是在特征空间进行的而不是在原始样本空间中计算, 其使用的距离公式不是通常的欧氏距离公式, 而是采用下式计算距离:

$$d(x,x_i)=\left|\phi(x)-\phi(x_i)\right|^2=k(x,x)-2k(x,x_i)+k(x_i,x_i),\ x_i\in T_{sv} \quad (3)$$

类似于 SVM, 针对不同应用问题可以选择式(3)中的核函数。算法中的分类阈值 ε通常设为 1 左右, 当 ε设为 0, KSVM 就是 SVM 算法。

4 实验结果与分析

4.1 文本分类模型的质量评估方法

文档分类中普遍使用的性能评估指标有查全率(Recall, 简记为 r)、查准率(Precision, 简记为 p)。二值分类的评估一般使用列联表。表 1 为一个二值分类问题的列联表。

表 1 二值分类问题的列联表

	真正属于该类的文档数	真正不属于该类的文档数
判断为属于该类的文档数	a	b
判断为不属于该类的文档数	c	d

这时, r 和 p 分别定义为: $r=\frac{a}{a+c}$ $p=\frac{a}{a+b}$

准确率和查全率反映了分类质量的两个不同方面, 两者必须综合考虑, 不可偏废, 因此, 存在一种新的评估指标: F1 测试值, 其数学公式如下:

$$F1=\frac{\text{准确率}\times\text{查全率}\times 2}{\text{准确率}+\text{查全率}}$$

4.2 选定 SVM 分类器最佳参数及分类器改进前后的实验结果比较

4.2.1 选定 SVM 分类器最佳参数设置

通过不断实验, 我们选定 SVM 分类器系统的最佳参数设置方案如表 2 所示。

表 2 系统最佳设定方案

分类器影响因素	特征词数量	训练集规模/每类	专利名称系数	特征词抽取比例	核函数选择	测试集规模/每类
设定值	150	300	5	1∶1	线性核函数	150

采用该系统最佳方案对测试集进行分类, 并且对其进行评估实验, 从实验结果可以看出, 经过选定各个最优设定, 系统的分类准确率和查全率都达到了接近 98%, 体现了支持向量机优秀的文本分类能力。

4.2.2 SVM 和 KSVM 两种分类器的实验结果比较

采用 SVM-KNN 算法进行分类也要用到分类器, 现在采用与 SVM 相同的多项式核函数。为了跟 SVM 比较, 同样采用了训练集每类样本数为 100, 实验结果如表 3 所示。然后采用最优方案对 SVM 和 KSVM 进行实验比较, 实验结果如表 4 和图 4 所示。

表 3 多项式核中 SVM 和 KSVM 的效果比较

参数		B21	C09	B25	D06	平均
		准确率/%				
q=1	SVM	95.8	92.2	96.6	97.3	95.5
	KSVM	96.5	95.4	96.9	97.0	96.5
q=2	SVM	97.7	85.8	93.6	100.0	94.3
	KSVM	96.3	95.8	95.5	96.4	96.0
q=3	SVM	100.0	72.5	91.9	100.0	91.1
	KSVM	96.2	95.2	94.8	96.4	95.7

表 4 采用最佳参数后两种算法的分类效果比较

评估值		A43	F02	B21	C09	B25	D06	E06	G02	平均
准确率	SVM	100.0	98.0	96.6	95.4	99.3	97.4	97.4	99.3	97.9
	KSVM	100.0	99.3	98.0	97.5	99.3	98.5	98.6	99.3	98.8
查全率	SVM	98.0	100.0	95.3	96.6	97.3	97.4	99.3	97.3	97.7
	KSVM	99.3	100.0	98.5	97.8	97.9	99.3	98.5	98.3	98.7
F1 值	SVM	99.0	99.0	95.9	96.0	98.3	97.4	98.3	98.3	97.8
	KSVM	99.7	99.7	98.3	97.7	98.6	98.9	98.6	98.8	98.8

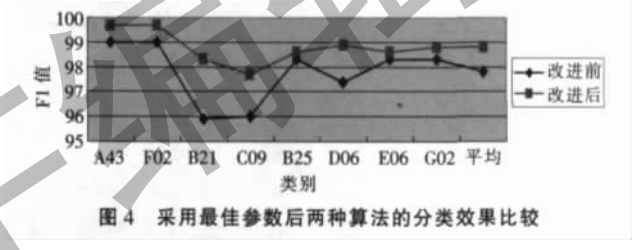

图 4 采用最佳参数后两种算法的分类效果比较

从实验的结果数据可以得出三个结论:

第一是使用 SVM-KNN 分类可以减轻对核函数参数选择的敏感程度, 缓解对参数选择的困难。对于 SVM 分类器, 核函数参数的选择是非常重要但很困难的。如表 3 中当参数 q=1、q=3 时, SVM 的分类性能差别很大。带入 KSVM 算法后, 对于参数的选择不是很敏感。如表 4 中的 q=1、q=2 和 q=3, KSVM 算法的效果差别很小, 性能比较稳定。

第二个结论是使用 SVM-KNN 分类器在一定程度上比使用 SVM 具有更好的性能。这可以从表 4 和图 4 看出来, 因为 KSVM 能很好地解决 SVM 对于超平面附近样本错识的问题, 因而明显的提高准确率。

第三是从表 3 和表 4 可以看出, 分类器采用了最佳参数后分类效果有了明显的提高, 这说明该专利分类器确实存在一组适合用于专利文本分类的一组参数, 这些参数可以通过公式推断评估、多次实验和经验确定下来。

5 结束语和未来的工作

从实验结果可以看出, 对于专利文本分类来说使用支持向量机技术是一个很好的方案。但目前在 SVM 的应用中还存在一些问题, 如对不同的应用问题核函数参数的选择较难, 对较复杂问题或超平面附近的向量其分类精度不是很高以及对大规模分类问题训练时间长等。因此系统采用了 SVM-KNN 组合算法对分类器进行改进, 并取得了很好的效果。

我们将进一步完善和加强专利文本分类系统, 使该系统能

(下转 212 页)

小，精度就越高。但这个算法也给服务器和客户端之间带来额外的数据流量，网格划分的越细，流量越大。

通过这种方式，服务器需要提供给客户端一组和目前客户端上显示地理方位相关的投影矩阵的参数，它要占据一定的数据通道。下面我们分析一下这些额外数据会比原来增加多少通讯流量。

假设图象的大小为 rows*cols，压缩率为 r（比如通过 jpeg 图像格式传送），三角网格的采样间隔为 gx，gy，参数大小为 s，那么：

图像流量 m=rows*cols*3*r；

额外流量 $n=\frac{6*s*(rows*2)}{gx*gy}$；

其中 s，r 在一个应用中都是固定的，假设格子的大小纵横比为 1，即 g=gx=gy：

$$q=\frac{n}{m}=\frac{4s}{g^2 r}$$

假定使用文本方式传递参数 s=10（10 个字符表示一个双精度的数字），图像压缩率为 30%，s=10，r=0.3，那么 $q=133.33/g^2$。

假设我们设置的格网 g=50，这时候额外的流量是 5.3%。

算法的精度和额外数据流量之间的具有较好的伸缩性，可以根据实际情况自由调节额外数据流量和精度比例。

这个算法还具有普遍性，可以适合任何投影，利用 PROJ4 等投影计算包可支持几十种投影，几百个椭球体。

6 实际软件结构

我们在现有的 WEB 地图发布引擎的基础上，可以很容易地加以改造。首先地图服务器获得一个地图的请求，在地图服务器生成图像的同时，转发给投影计算服务器（PROJ4 投影计算包）根据地图使用的投影方式和已经和客户端协定好的三角网划分方案，计算出各个定点的地理坐标系位置和经纬坐标系位置，在它们之间建立三角插值参数。通过网络把地图影像和插值参数一起发送到客户端，客户端上有一个简单的三角网插值函数把捕获的客户端鼠标事件的位置代入求出经纬度数值，然后动态显示在界面上。

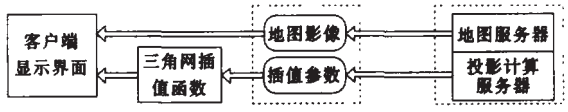

图 2 软件结构

在实际应用中有些客户端可能没有浮点计算能力（比如一些手机上的 Java 平台），三角网插值也可以很容易地用定点计算模式来实现，因此客户端的适用面非常广。

7 结论

在 WebGIS 上动态显示经纬度坐标是很重要的功能，但是如果显示的数据集本身不是经纬度坐标系的，那么就涉及投影转换的问题。

直接在客户端上做各种地理投影的算法实现是不现实的，目前的各种商业产品都没有很好的解决办法。

本文的方法是通过线性逼近复杂函数的原理，把计算分解成两个部分，在服务器做复杂的投影计算，在客户端做简单的插值计算，实现了这个功能。这个设计的优点在于它的伸缩性和普遍性，有很好的应用前景。（收稿日期：2005 年 10 月）

参考文献

1.郭仁忠.空间分析[M].武汉测绘科技大学出版社，2000
2.杨启和.地图投影变换原理与方法[M].北京：解放军出版社，1989
3.李国藻，杨启和，胡定荃.地图投影[M].北京：解放军出版社，1993
4.钱曾波，刘静宇，肖国超.航天摄影测量[M].北京：解放军出版社，1992
5.任留成.空间投影理论及其应用研究[D].郑州：解放军测绘学院，1999
6.易大义，陈道琦编.数值分析引论[M].浙江大学出版社，1998
7.任留成，杨晓梅.空间 Gauss- Kruger 投影研究[J].测绘学院学报，2004；21(1)：73~75
8.滕骏华，孙美仙，黄韦艮.地图投影反解变换的一种新方法[J].测绘学报，2004；33(2)：179~185
9.GCTPC.http：//edcftp.cr.usgs.gov/pub//software/gctpc/
10.PROJ4.http：//proj.maptools.org/

（上接 97 页）

169~174
2.郭关飞，张尧学，周悦芝.透明计算：可管理多媒体网络计算机[J].高技术通讯，2004；14(增刊)：57~61
3.张尧学，彭玉坤，周悦芝等.可管理多媒体网络计算机(MMNC)[J].电子学报，2003；31(12A)：2054~2058
4.张尧学，周悦芝，王勇等.一种网络环境下的计算机远程启动方法[P].中国：发明专利，申请号：01142033.2
5.周悦芝，张尧学，王勇.一种用于网络计算的可定制启动协议[J].软件学报，2003；14(3)：538~546
6.M E Russionovich，D A.Solomon.Windows Internals[M].4th edition，Washington，USA：Microsoft Press，2004：251~288
7.毛德操，胡希明.Linux 内核源代码情景分析(下册)[M].浙江大学出版社，2001：663~746

（上接 195 页）

够更好的融合到专利战略分析系统中去，进一步研究 KNN 及其他算法，以提高专利分类的分类效果。

（收稿日期：2006 年 1 月）

参考文献

1.李淑文.试论文本自动分类[J].现代计算机，2004；(7)
2.柳回春，马树元.支持向量机的研究现状[J].中国图像图形学报，2002；(6)
3.边肇祺，张学工等.模式识别[M].北京：清华大学出版社，2000：286~294
4.张学工.关于统计学习理论与支持向量机[J].自动化学报，2000；26(1)：32~43
5.王国胜，钟义信.支持向量机的若干新进展[J].电子学报，2001；(10)
6.Vapnik V N.The Nature of Statistical Learning Theory[M].NY：Springer Verlag，1995
7.李蓉，叶世伟，史忠植.SVM- KNN 分类器——一种提高 SVM 分类精度的新方法[J].电子学报，2002；30(5)：745~748
8.李红莲，王春花，袁保宗.一种改进的支持向量机 NN- SVM[J].计算机学报，2003；26(8)：1015~1020