SVM-*K*NN 分类器——一种提高 SVM 分类精度的新方法

李 蓉, 叶世伟, 史忠植
(1 中国科技大学研究生院(北京)计算机教学部, 北京 100039; 2. 中国科学院计算技术研究所智能信息处理实验室, 北京 100080)

摘 要: 本文提出了一种将支持向量机分类和最近邻分类相结合的方法, 形成了一种新的分类器. 首先对支持向量机进行分析可以看出它作为分类器实际相当于每类只选一个代表点的最近邻分类器, 同时在对支持向量机分类时出错样本点的分布进行研究的基础上, 在分类阶段计算待识别样本和最优分类超平面的距离, 如果距离差大于给定阈值直接应用支持向量机分类, 否则代入以每类的所有的支持向量作为代表点的 *K* 近邻分类. 数值实验证明了使用支持向量机结合最近邻分类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.

关键词: 支持向量机; 最近邻分类; 类代表点; 核函数; 特征空间; VC 维
中图分类号: TP301 文献标识码: A 文章编号: 0372-2112 (2002) 05-0745-04

SVM-*K*NN Classifier——A New Method of Improving the Accuracy of SVM Classifier

LI Rong, YE Shi-wei, SHI Zhong-zhi
(1. *Dept. of Computing, Graduate School, Science and Technology University of China, Beijing* 100039, *China*; 2 *National Key Laboratory of Intelligent Information Technology Process, The Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: A new algorithm that combined Support Vector Machine (SVM) with *K* Nearest neighbour (*K*NN) is presented and it comes into being a new classifier. The classifier based on taking SVM as a 1NN classifier in which only one representative point is selected for each class. In the class phase, the algorithm computes the distance from the test sample to the optimal super plane of SVM in feature space. If the distance is greater than the given threshold, the test sample would be classified on SVM; otherwise, the *K*NN algorithm will be used. In *K*NN algorithm, we select every support vector as representative point and compare the distance between the testing sample and every support vector. The testing sample can be classed by finding the k-nearest neighbour of testing sample. The numerical experiments show that the mixed algorithm can not only improve the accuracy compared to sole SVM, but also better solve the problem of selecting the parameter of kernel function for SVM.

Key words: support vector machine; nearst neighbour algorithm; representative point; kernel function; feature space; VC Dimension

1 引言

统计学习理论是一种专门的小样本统计理论, 为研究有限样本情况下的统计模式识别和更广泛的机器学习问题建立了一个较好的理论框架, 同时也发展了一种模式识别方法—支持向量机(Support Vector Machine, 简称 SVM), 在解决小样本、非线形及高维模式识别问题中表现出许多特有的优势, 并能够推广应用到函数拟合等其他机器学习问题中[1]. 目前, 统计学习理论和SVM 已经成为国际上机器学习领域新的研究热点并已被应用于人脸识别、文本识别、手写体识别等领域.

在对 SVM 的研究中, 提高它的分类能力(泛化能力)是所有研究的出发点和归宿. SVM 和其他分类方法相比具有较高的分类精度, 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[1] 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.

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[1], 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的样本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(详细推导见附录). 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一给定的阈值, 即样本离分界面较远, 则用 SVM 分类, 反之用 *K*NN 对测试样本分类. 在使用 *K*NN 时以每类的所有的支持向量作为

收稿日期: 2001-06-15; 修回日期: 2001-10-06

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.

2 SVM、*K*NN 分类器简介

2.1 SVM 分类器

SVM 是一种建立在统计学习理论基础上的分类方法[1]. 它主要基于以下三种考虑(1)基于结构风险最小化, 通过最小化函数集的 VC 维来控制学习机器的结构风险, 使其具有较强的推广能力. (2)通过最大化分类间隔(寻找最优分类超平面)来实现对 VC 维的控制, 这是由统计学习理论的相关定理保证的. (3)而 SVM 在技术上采用核化技术, 根据泛函中的 Mercer 定理, 寻找一个函数(称核函数)将样本空间中内积对应于变换空间中的内积, 即避免求非线形映射而求内积.

2.2 *K*NN 分类器

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"[11]. 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, 结果就是与 x 最近的训练样本所属于的类别. *K*NN 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.

3 SVM-*K*NN 分类器实现

3.1 对 SVM 分类机理的分析

在本文中, 通过对 SVM 的分类机理分析, 找到了 SVM 和 NN 分类器之间的联系, 此联系由下面的定理给出:

定理 1 SVM 分类器等价于每类只选一个代表点的 1-NN 分类器.

证明见附录.

3.2 SVM-*K*NN 分类器的形成

将 SVM 和 *K*NN 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 *K*NN 相结合是因为 *K*NN 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, 如果距离差大于一给定的阈值即 x 离分界面较远, 如图 1 中区域 Ⅱ, 用 SVM 分类一般都可以分对. 当距离差小于一给定的阈值, 即 x 离分界面较近, 即落入区域Ⅰ, 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 *K*NN 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.

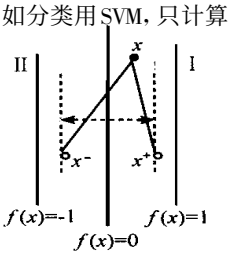

图 1

在对数据的封闭测试(用训练样本作为测试集)中, SVM-*K*NN 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Ⅰ, 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. 正由于上述原因, 在下面的实验中不进行封闭测试.

3.3 SVM-*K*NN 分类器算法(*K*SVM 算法)

*K*SVM 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, T_{sv} 为支持向量集, k 为 *K*NN 的个数.

Step1: 如果 $T\neq\Phi$, 取 $x\in T$, 如果 $T=\Phi$, 停止;

Step2: 计算公式 $g(x)=\sum_i y_i\alpha_i K(x_i, x)-b$.

Step3: 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 *K*NN 算法分类, 传递参数 x、T_{sv}、k, 返回结果为输出.

Step4: $T\leftarrow T-\{x\}$, go to Step1.

上述算法 Step3 中使用的 *K*NN 分类算法和通常的 *K*NN 分类算法流程相同, 将支持向量集 T_{sv} 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.

$$d(x, x_i)\|\phi(x)-\phi(x_i)\|^2=k(x, x)-2k(x, x_i)+k(x_i, x_i),\quad x_i\in T_{sv} \qquad (1)$$

类似于 SVM, 针对不同应用问题可以选择式(1)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, *K*SVM 就是 SVM 算法.

4 实验结果及分析

进行两组实验, 第一组是对经典的模式识别问题双螺旋线的实验, 第二组是对文本分类的实验. 选择这两组实验的目的是比较对于不同应用问题 SVM 和 KSVM 分类算法的性能. 实验方法采用 Bootstrap 方法: 对于 n 个样本的数据集随机抽取 n 次, 将取到的样本去掉重复作为训练样本, 余下的样本作为测试样本. 这种抽样方法所得到的训练样本集约占总数据集的 63.2% . 重复上述操作十次, 取平均结果.

(1)双螺旋线测试

双螺旋线问题属于两类分类问题, 如图 2 所示. 可将图中粗点表示的螺旋线数据作为正例, 细点表示的螺旋线数据作为反例. 螺旋线圈数的多少代表了问题的不同复杂程度的问题. 圈数越多问题可分性越差, 分别对两圈、三圈和四圈的螺旋线进行实验, 所选择的核函数为高斯核函数 $k(x, x_i)=\exp(-g^*\|x-x_i\|^2)$, 错误

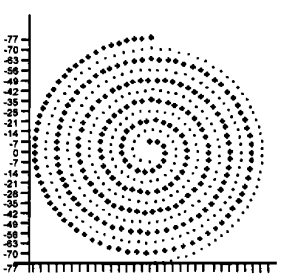
图 2 双螺旋线图示

惩罚参数 $C=5$, 分类阈值 ε 选为 0.8. 选择了四组不同的核函数参数测试, 比较两种算法对不同参数的分类效果. 实验结果如表 1 所示.

(2) 文本分类实验:

将下载的 5642 个中文网页后通过人工方式将其分为十三类后, 对各个类标明其输出. 这时一个多类分类问题, 针对此多类问题我们构造了 SVM 多值分类器, 构造方法采取一对一方式[4], 训练了 $\frac{k^*(k+1)}{2}$ ($k=13$) 个 SVM 二值子分类器. 本次实验所选取的核函数为感知机核函数 $k(x, x_i)=\tanh(g^*(x^* x_i)+c)$, 大量数据测试证明对于网页分类数据, 采用感知机核函数在分类准确率和速度上均优于其它核函数. 在此实验中错误惩罚参数 $C=5$, 分类阈值 ε 取为 0.6. 除了对综合测试集进行测试外, 我们还从中选取了有代表性几个类分别测试, 测试结果如表 2 所示.

表 1　双螺旋线分类 SVM 和 KSVM 算法比较

核参数	分类算法	圈数: 2	圈数: 3	圈数: 4
$g=0.5$	SVM	54.7312%	50.9241%	47.1546%
	KSVM	49.3677%	48.4618%	50.0917%
$g=0.05$	SVM	61.6282%	50.9241%	50.6731%
	KSVM	95.7631%	86.3446%	81.0137%
$g=0.03$	SVM	81.6002%	82.1874%	72.8237%
	KSVM	92.8041%	86.3446%	85.1858%
$g=0.01$	SVM	95.9519%	87.8010%	57.6668%
	KSVM	95.7631%	86.3446%	85.1876%

表 2　对于文本分类 SVM 和 KSVM 算法比较

核参数	分类算法	综合类	工业类	体育类	生活类	政治类
$g=2$	SVM	65.1423%	56.9759%	83.8684%	63.3834%	75.7044%
	KSVM	68.8713%	60.3927%	88.8192%	64.5993%	78.3995%
$g=0.5$	SVM	66.6612%	59.888%	83.3060%	66.4731%	81.4176%
	KSVM	69.1269%	62.0845%	87.9798%	65.5740%	82.2401%
$g=0.1$	SVM	46.2187%	2.9668%	59.4340%	26.8909%	87.9119%
	KSVM	64.1182%	61.8701%	85.3217%	54.3182%	89.1481%
$g=0.05$	SVM	30.2999%	0%	31.3306%	0%	92.7028%
	KSVM	64.0689%	61.3808%	82.9425%	51.1887%	93.9405%

(3) 实验分析

从实验的结果数据可以得出两个结论: 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 1 中当参数 $g=0.5$、$g=0.01$ 及表 2 中的 $g=0.5$、$g=0.05$, SVM 的分类性能差别很大. 对于同一参数, 问题不同分类效果差别也很大, 如上表 1 中 $g=0.01$, 对圈数为二、三的螺旋线, SVM 的分类效果很好, 但对于四圈的螺旋线, SVM 的识别率不如选择 $g=0.03$ 的识别率. 带入 KSVM 算法后, 对于参数的选择不是很敏感. 如表 1 中的 $g=0.05$ 和 $g=0.01$, 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. 针对四圈情况, 数据的线形不可分程度高, 使用 SVM 分类性能低, 而使用 KSVM 算法分类精度提高较明显. 而当实际问题相对好分时(表 1 中的二、三圈螺旋线), 二者的分类效果差别不大. 这是因为当实际问题比较易分时, SVM 训练后所得到支持向量个数少, 在 KSVM 中所选取的代表点也少; 实际问题复杂程度高时, SVM 训练后所得到支持向量个数多, KSVM 算法所携带的信息更高, 而此时 SVM 分别对正反例支持向量组仅合成一个代表点, 损失的信息也相对较多.

5　结论

本文将 SVM 和 KNN 分类器相结合构造了一种新的分类器, 该分类器基于将 SVM 分类器等价于对每类支持向量只取一个代表点的 1NN 分类器, 针对当测试样本在分界面附近时容易分错的情形引入 KNN 分类选取每个支持向量作为代表点. SVM-KNN 分类器是一种通用的分类算法, 使用该分类器的分类准确率比单纯使用 SVM 分类器一般有不同程度的提高, 并且该分类器在一定程度上不受核函数参数选择的影响, 具有一定的稳健性. 进一步的工作是从 SVM 的分类机理得到启发, 不一定采用每个支持向量作为代表点, 而对它们进行组合选取多个代表点进行分类.

附录: 定理 1 证明

已知线性可分样本集为 (x_i, y_i), $i=1, \cdots, l$, $x_i \in R^d$, $y \in \{-1, +1\}$ 为类别标志, d 空间维数. 最优分类面问题可以表示成如下优化问题的对偶问题

$$W(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(x_i^* x_j) \tag{1}$$

约束条件为:

$$0\leq\alpha_i,\ i=1, \cdots, n \text{ 与 } \sum_{i=1}^{l}\alpha_i y_i=0 \tag{2}$$

根据 Kuhn-Tucker 条件, 这个优化问题的解必须满足

$$\alpha_i(y_i[(w, x_i)-b]-1)=0,\ i=1, \cdots, l \tag{3}$$

相应的分类函数也变为

$$f(x)=\operatorname{sgn}\left(\sum_i y_i\alpha_i^* k(x_i, x)-b\right) \tag{4}$$

首先分别利用正、反例支持向量组成两个代表点 $\phi(x)^+=\frac{1}{C}\sum_{y_i=1, i=1}^{l}\alpha_i\phi(x_i)$, $\phi(x)^-=\frac{1}{C}\sum_{y_i=-1, i=1}^{l}\alpha_i\phi(x_i)$. 其中 $\sum_{y_i=1}\alpha_i=\sum_{y_i=-1}\alpha_i=C$ (根据目标函数对偶问题的等式约束条件 $\sum_{i=1}^{l}\alpha_i y_i=0$), 对于最优解 $w=\sum_{i=1}^{l}\alpha_i\phi(x_i)=C(\phi(x)^+-\phi(x)^-)$, 由式(3)对任意正例的样本有 $\alpha_i((w, \phi(x_i))-b-1)=0$, 从而有

$$\begin{aligned}0&=\sum_{y_i=1}\alpha_i((w, \phi(x_i))-b-1)\\&=(w, \sum_{y_i=1}\alpha_i\phi(x_i))-C^* b-C\\&=(C(\phi(x)^+-\phi(x)^-), C\phi(x)^+)-C^* b-C\\&=C[C((\phi(x)^+-\phi(x)^-, \phi(x)^+))-b-1]\end{aligned} \tag{5}$$

这样有 $b=C(\phi(x)^+-\phi(x)^-, \phi(x)^+)-1$

同样由式(3), 对任意反例的样本有

$$b = C((\phi(x)^+ - \phi(x)^-, \phi(x)^-)) + 1 \quad (6)$$

由(式(5)+式(6))/2可得

$$b = \frac{C}{2}((\phi(x)^+ - \phi(x)^-, \phi(x)^+ + \phi(x)^-))$$

$$= \frac{C}{2}(k(x^+, x^+) - k(x^-, x^-)) \quad (7)$$

在SVM的分类过程代入1NN分类,可得到下式:

$$g(x) = \|\phi(x) - \phi(x)^-\|^2 - \|\phi(x) - \phi(x)^+\|^2$$

$$= 2k(x, x^+) - 2k(x, x^-) + k(x^-, x^-) - k(x^+, x^+)$$

$$= \frac{2}{C}\left\{\sum_i \alpha_i y_i k(x, x_i) + \frac{C}{2}[k(x^-, x^-) - k(x^+, x^+)]\right\}$$

(由式(7)可得)$$= \frac{2}{C}\left\{\sum_i \alpha_i y_i k(x, x_i) - b\right\} \quad (8)$$

参考文献:

[1] Vapnik V N. The Nature of Statistical Learning Theory [M]. NY: Springer Verlag.

[2] 边肇祺, 张学工. 模式识别[M]. 北京: 清华大学出版社. 2000.

[3] Vapnik V N. Estimation of dependencies based on empirical data [R]. Berlin: Springer Verlag, 1982.

[4] Burges C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discvery, 1998, 2(2).

[5] Weston J, Watkins C. Multi-class support vector [J]. machines. Royal Holloway College, Tech Rep: CSK-TR-98-04, 1998.

[6] Thorston Joachims. Text Categorization With Support Vector Machine: learning with relevant features [R]. University Dortmund, 1998.

作者简介:

李 蓉 女, 1973 年生于北京, 1996 年于北京理工大学获工学学士位, 1999 年进入中科技大学研究生院(北京)计算机软件专业攻读硕士学位, 2000 年 10 月至今在中科院计算技术研究所智能信息处理开放实验室做硕士论文, 师从于史忠植研究员, 研究方向为机器学习、神经计算, 已发表学术论文 3 篇.

叶世伟 男, 1968 年生于四川, 分别于 1991 年、1993 年、1996 年于四川师范大学、北京大学、中科院计算技术研究所获得理学学士、理学硕士和工学博士学位, 现任中科院研究生计算机教学部院副教授, 主要研究方向为神经计算、优化理论, 已发表学术论文十余篇.

2002 国际存储展览暨技术研讨会在京召开

由信息产业部电子信息产品管理司、中国电信、国家邮政局及中国计算机学会信息存储技术专业委员会支持, 中国电子信息产业发展研究院(CCID)主办, 赛迪展览公司承办的"2002 国际存储展览暨技术研讨会(Storage Infoworld 2002)"4 月 25~27 日在北京中国国际科技会展中心隆重举行. 信息产业部苟仲文副部长参加开幕主题演讲并致欢迎辞, 随后在信息产业部有关司局领导的陪同下饶有兴趣地参观了展览会, 并与参展企业代表亲切座谈. 来自各有关部委和行业用户部门的三十多位领导和近千余名专业人士出席了展览及研讨会.

Sotrage Infoworld 2002 聚焦存储领域热点, 汇聚如 EMC、SUN、HP、Network Appliance、Xiotech、Seagate、CA、Auspex、RC、Spectra Logic、VERITAS、Quantum、Maxtor、SONY、ANEKtek、清华同方、亚美联等三十余家国内外知名存储软硬件厂商、存储系统集成商、存储技术开发商及相关的经销商和渠道合作伙伴, 内容涵盖网络存储、光存储、移动存储、存储软件及存储应用解决方案. EMC 公司在展会上推出了一系列高级、整合并经过验证的业务连续性解决方案; Sun 公司的 Storage ONE 体系架构提供了一个开放、集成化和自动的存储管理解决方案; Network Appliance 作为数据存储和内容传输领域的世界领先者, 为当今数据激增的企业提供开放的网络存储解决方案; 亚美联公司作为国内首家完全自主知识产权的企业级存储产品供应商, 推出的国内第一套达到国际先进技术水平的企业级存储系统 Estor NAS18/2800、Estor SAN 产品系列; Spectra Logic 公司的 Spectra 64000 企业级磁带库、昆腾公司的基于磁盘的产品系列——第一款产品 Quantum DX30 等都在展会上亮相.

在两天半的研讨会中, 来自 EMC、SUN、HP、XIOtech、CA、Spectra Logic、清华同方等公司的国内外存储专家, 将从存储的最新动态、发展方向、最新技术、解决方案和成功案例等方面发表精彩演讲.

IT 界称 2001 为存储年, 而 2002 年将为中国存储市场迎来又一高峰. Storage Infoworld 2002 作为国内 IT 存储领域举办的权威盛会, 必将以优质的服务为国内外关注中国存储市场发展的厂商及用户架起供需沟通的桥梁.