核方法及其在 KCPA 及 SVM 中的利用

王尧　无 54　2005011068

[摘要] 首先介绍了近年来人工智能领域广泛关注的研究热点——核方法的基本思想，分析了使用核函数给我们的研究带来了怎样的好处。进而列举了3类重要的不同核函数，并对其不同的性质做了简单的比较。最后结合实际案例，分别在SVM（支持向量机）及KCPA（核主成分分析）中使用不同的核函数进行手写数字的识别，其中SVM主要研究了多种核函数的效率及其有效性；而在KPCA中则进行了再生核函数性质的试探分析，并提出了一种选择高效核函数的试行方法——选择适当阶次的再生核函数能有效地提高识别效率，获得了预计的效果。

[关键词]

核函数　再生核函数　SVM　KPCA　比较　手写体识别试验

[核函数介绍]

1-1 核函数由来

核方法顾名思义就是利用核函数来进行研究的方法，其对非线性分类的解决提供了独特快速的支持，而现实中很多分类识别问题都是非线性的，因此基于核函数的研究方法备受大家关注，成为研究的热点。对于非线性分类，首先使用一非线性映射函数$\phi(x)$把数据从原空间$\mathbf{R}^d$ 映射到一个高维特征空间Ω,再在高维特征空间Ω建立优化超平面进行问题的处理，考虑到在线性情况处理时一般只用到了原空间的内积运算，在非线性空间也只　　考虑在高维特征空间Ω的内积运算$<\phi(x), \phi(y)>=K(x,y)$ ，这里的K（x,y）称为核函数。核函数本质上是一个内积，即通过引入核函数，把基于内积运算的线性算法非线性化，让我们可以不去考虑$\phi(x)$的非线性变化过程，这样就节省大量的时间，提高了研究效率。

1-2 核函数及基本性质

1-2-1核函数定义　称二元函数$k: X \times X \to R$是核函数，如果存在某个内积空间$(H, \langle \cdot, \cdot \rangle)$，以及映射$X \to H$使得$K(x,y) = <\phi(x), \phi(y)>$。（称H为特称空间，$\phi$为特征映射。

1-2-2正定性：　$k: X \times X \to R$是核函数当且仅当它是正定的。

1-2-3封闭性：若k1,k2,…是核函数，则

（1）　k1+k2是核函数。

（2）　ak1, a>0是核函数。

（3）　k1*k2是核函数。

1-2-4　设$k: X \times X \to R$是核函数，$f: X \to R$是任意函数，则：

（1）　$f(x)f(x')$是核函数。

（2）　$f(x)k(x,x')f(x')$是核函数。

把一些已知的核函数作为基本模块，根据以上的一些性质，便能构造出许多新核函数。这为之后的研究不同核函数的不同效果，及其研究不同问题怎样选择合适的核函数提供了一

定的帮助。

1-3 核函数反映了输入数据之间的相似性

设 k：X × X →R 是核函数，ϕ 是 k 的特征映射，则 k 在输入空间 X 上诱导了一个伪距离：

$$\begin{aligned}\rho_k(x,x') &= \| \Phi(x) - \Phi(x') \| \\ &= \sqrt{\langle \Phi(x),\Phi(x)\rangle - 2\langle \Phi(x),\Phi(x')\rangle + \langle \Phi(x'),\Phi(x')\rangle} \\ &= \sqrt{k(x,x) - 2k(x,x') + k(x',x')}\end{aligned}$$

这个伪距离可解释成x与x′之间的相似性度量，这正是核方法的精髓关键所在。正是由于核函数的这个性质，我们将之广泛了运用到了识别分析中。广义上来讲，我们可以将核函数定义到更抽象复杂的空间上。比如声音、图片、字符串。这样我们就能利用核函数进行具体的相似性识别。这样怎样提取物体的特性，构造对应的核函数就是我们需要解决的问题，后文将以手写数字的识别作为示例来进行分析。

1-4 3类重要核函数

（1） 平移不变核

平移不变核是指核函数具有形式 k(x,x′)=f(x-x′)，其中 f()是实函数。例如，

高斯核函数 $k(x,x') = e^{-a\|x-x'\|^2}, a>0$

指数径向基核 $k(x,x') = e^{-a\|x-x'\|}, a>0$

（2） 旋转不变核

旋转不变核指核函数具有形式k(x,x′)=f(<x,x′>),其中f()是一元实函数。例如，

齐次多项式核函数 $k(x,x') = \langle x,x'\rangle^p, p\in\mathbb{N}$ 。

非齐次多项式核函数 $k(x,x') = (\langle x,x'\rangle + 1)^p, p\in\mathbb{N}$

感知器核函数 $k(x,x') = \tanh(\rho\langle x,x'\rangle + c), \rho, c>0$

（3） 卷积核

这也是一种常用的核函数构造方法，即两个核函数的卷积也是一个核函数，这里借用了信号与系统的卷积概念，二者及其相似，结合核函数的频域来考虑能够有效地获得其卷积核函数，即再生核函数。

掌握这 3 类核函数的特点对我们用核方法研究问题有着重要的帮助，获得较好的算法效果和选择适当的核函数是分不开的。

[研究方法]

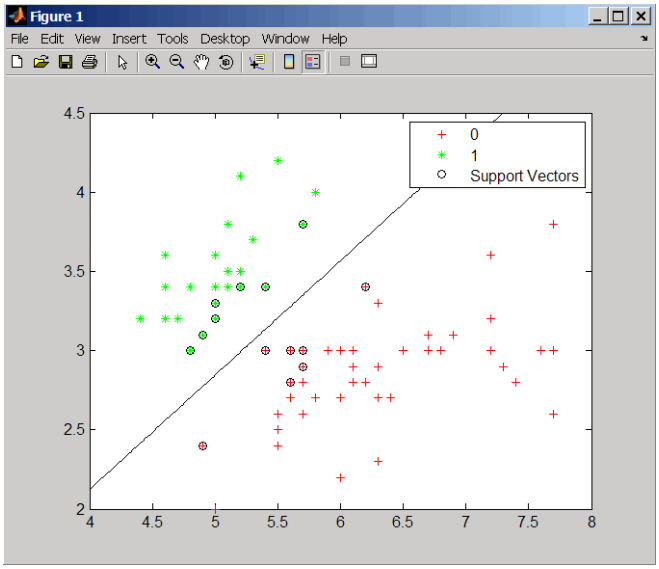
在学习分析了核函数的以上知识和性质和，我们就要将它与具体的分析算法结合起来来解决具体的实际问题。我先后分别选用了比较热门的SVM（支持向量机）和KPCA（核主成分分析）进行研究。

2-1 SVM（支持向量机）

支持向量机是核函数的重要使用领域，它是在统计学习理论基础上发展的一种新的通用的学习方法，目前已表现出很多优于已有方法的性能。是大家研究的热点之一

2-1-1. SVM 基本原理

支持向量机首先是从线性可分情况下的最优分类面发展而来的，基本思想可用二维情况说明，如下图：



图中的绿点和红点分别代表两类不同的训练样本，中间实线为 2 者的分界线。两类样本点到分界线的最小距离之和称为分类间隔(margin).所谓最优分类线就是要求分类线不但能将两类正确分开即使训练错误为 0，而且要求分类间隔最大

假定大小为 a 的训练样本集{(xi,yi),i= 1,2,. ..a}，由二类别组成:若 xi 属于第 1 类，标记为(yi=1) ；若属于第 2 类，则标记为(yi=-1)。机器学习的目标是构造一判别函数，将测试数据尽可能正确地分类。

对于训练样本在线性可分情况下就会存在一个分类超平面 xw + b = 0 进行归范化使得线性可分样本集满足:

$$yi[(w^* xi) +b]\geq 1 , (i=1...,a) \qquad \text{————（A）}$$

由统计学习理论知，最优超平面就是指训练样本集没有被超平面错误分开，并且距超平面最近的样本数据与超平面之间的距离最大，由此得到的判别函数:f(x)= sgn ((w * x) +b)其泛化能力最优，sgn(.)为符号函数，此时分类间隔等于 2/||w|| ，使间隔最大等价使||w||最小。使分类间隔最大实际上就是对推广能力的控制，这是 SVM 的核心思想之一。

当训练样本集线性不可分时，条件(A)需引入松驰变量 ζi≥0 ,i= 1,2,...,a，此时(A)式变为

$$yi((w^*xi)+b]+ \zeta i\geq 1 , i=i,..,a \qquad \text{————(B)}$$

将目标改为求(w,ζi)=1/2*||w||^2+C$[\sum_{i=1}^{a} \zeta i]$最小，从而综合考虑最少错分样本和最大分类间隔，就得到广义最优分类面。其中，C 是惩罚因子，控制对错分样本惩罚的程度。利用 Lagrange 优化方法可以把上述最优分类面问题转化为其对偶问题。

训练样本集为非线性时，可以将训练样本集通过非线性变换转化为某个高维空间中的线性问题，在变换空间中构造最优分类面。但在求解毓化问题和计算判别函数时并不需要显式计算

该非线性函数，而只需计算核函数。根据泛函的有关理论，只要一种核函数 K(x,y)，满足 Mercer 条件，它就对应某一变换空间中的内积。因此，在最优分类面中采用适当的内积核函数 K(x,y)，就可以实现非线性变换后的线性分类，而计算复杂度却没有增加

2-2 KPCA（核主成分分析）

KPCA 是将多年来应用较多的统计方法——PCA（主成分分析）与核函数结合起来使用的一种方法。PCA 充分利用数据中的二阶统计量对数据进行特征提取和降维处理，其主成分(PC)对应于数据均方误差重建误差曲面的最小点，因此对数据信息具有较强的描述能力，并且算法简单、运算量小。KPCA 则是先对数据进行升维处理，利用核函数对数据做非线性变换映射到高维空间中，然后利用 PCA 的研究方法对其进行分析。这样可以有效地解决低维空间中出现的非线性不可分问题。

PCA 的作用原理这里就不赘述了，之后的实验中将会介绍具体的实现过程。

[具体实验]

这里我们进行了对手写体数字的识别研究，使用 matlab 作为实验工具，分别采用了 svm 与 kpca 方法。二者在识别分析环节有些差距，但在初期的数据采集、处理和特征提取，及之后的效率分析等模块都是共同的。而且 matlab 具有专门的 svm 工具箱，直接调用其 svm 函数对其进行参数设置就可以进行 svm 分析，十分方便，而不用额外花大量时间。所以主要进行了 KPCA 模块的编程实现。

实验步骤：

（1） 数据读取　本次实验手写体数据均来自 http://yann.lecun.com/exdb/**mnist**/ 提供的数据，根据网站的使用说明可以清楚地知道字体数据的储存结构，使用 matlab 读文件即可。

（2） 数据分析及处理　mnist 数据库将每个手写体数字图片分为 28*28 个象素，以一个 28*28 的矩阵表示，分别储存 28*28 个象素的灰度值，我们可以将其转化为一个 1*784 的行向量作为一附图的特性表征。但在图形读取输出显示后分析发现，由于手写数字位于图片中心区域，周围的象素点几本均为 0。如下上图所示

（28*28 的矩阵表示的数字）

（22*22 的矩阵表示的数字）

为了增大不同数字的差异性，我选择切除边缘的一圈相似层，即仅将 28*28 矩阵的中心 22*22 矩阵作为我们的数据分析处理对象（如上下图所示），这样不仅可以缩小我们的数据处理规模，减小映射的特征空间的维度和部分步骤的时间复杂度及空间复杂度，还有助于微量的提升数据的识别正确率。实验证明，切变处理前后处理率

平均增加了 0.08，且识别时间有了较大的减少。

	28*28 点数据	22*22 点数据
识别率	83.5%	91.5%
识别时间	19.6 秒	14.2 秒

（以指数径向基函数，识别 200 组数据为例，采用 KPCA 处理方法）

为了增加识别精度，对象素灰度矩阵进一步进行了 2 值处理，即灰度值高于 1 则置其为 1，低于中值则置其为零。然后对其进行滤波处理，以消除 2 值矩阵中的孤立点，减少数据噪声。

经过如上处理后，每个字体图片就对应了一个 1*484 的 2 值行向量。然后我们就可以对这些向量们代替图片进行识别操作，寻找他们的特性与共性。

（3） 识别处理

svm：

由于 SVM 是基于 2 类样本点的区分的。Matlab 里面也有对应的工具箱。只需人工另外编写设置核函数。所以我选择使用已有函数来实验的做法。为了用两类分类机制来实现 0-9 十个数字的识别，我的基本思想就是采用逐个多次识别的方法。即分别将样本点分为 0 与非 0，1 与非 1，……，9 与非 9。这样分别进行机器学习，得到 10 组识别结构体。然后将未知样本分别与这 10 组识别结构体进行判别比较。最后进行综合分析处理，即可得到最终识别分类结果。

首先将从 minst 的读出的图片数据向量组和其真值数组一一对应，调用 svmtrain 函数对其进行训练。其中的 Kernel_FunctionValue（核函数模型）参数自己手动设定。然后对未知的数据组我们调用 svmclassify 函数对其进行分类处理即可。

- **不同的核函数的识别效果是不同的** 这个道理显而易见，不同的映射方法必然导致分类效果的区别分别采用了线性内积、高斯核函数、指数径向基核、齐次多项式核函数、非齐次多项式核函数对其进行识别，识别率其使用时间分别如下图所示。

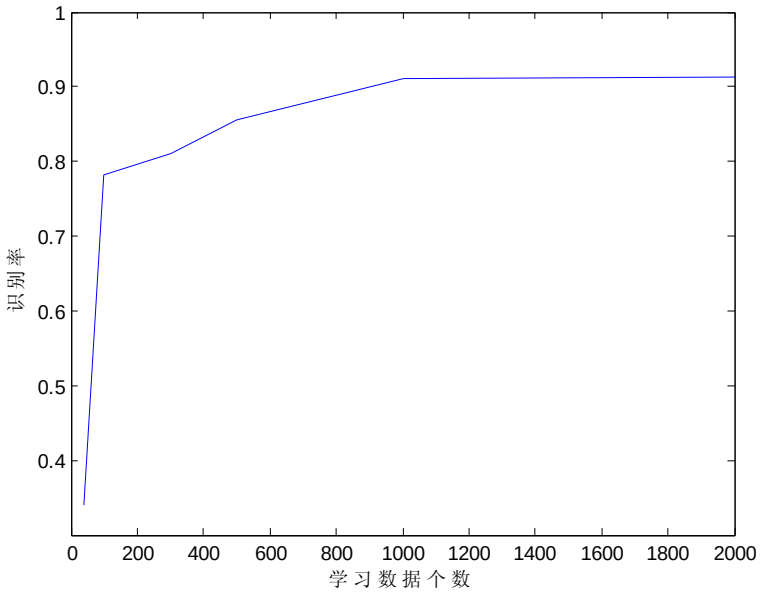
核函数	识别率
线性内积	0.35
高斯核函数	0.88
指数径向基核	0.915
齐次多项式核	0.8
菲齐次多项式核	0.8

- **同一种核函数的参数选择不同也会给识别带来一定的影响**，以径向基函数为例

δ ^2	1/2	1/4	1/8	1/12	1/16	1/18	1/20
识别率	0.47	0.68	0.79	0.85	0.932	0.92	0.83

参数的选择给识别率带来一定影响，合理的选择适当的识别率有利于识别率的提高。但是目前仅仅能采用试凑法，比较具有随机性，至于对特定的核函数我们该具体如何选用参数则是我们需要继续研究的问题。

•训练样本的个数不同也是影响 SVM 识别率的一大关键因素，只有保证机器进行一定数目的样本学习后，才能达到理想的识别效果



小结：

svm 中训练样本个数的选择对最终的识别率有着较大的影响，机器学习更多的训练样本得到识别参数更加精确，更能体现不同样本空间的差异，因而能够获得更高的识别率。不同的核函数的选择也会导致识别率的不同。同类核函数的参数选择不同也会导致识别率的高低。所以选择合理的核函数并设置适当的参数是十分重要的。

KPCA：

用核主成分分析法我们必须首先对所有的 10 类样本分别构造核矩阵，来表征低维空间到高位空间的映射关系。即原理正是基于核函数的性质。

首先将不同学习样本的数据进行分类，对每个数字分别收集N个数据（xi,i=1,…,N）作为核矩阵构造基础。然后根据自己定义的核函数构造出一个N*N的核矩阵K，其中Kij=K(xi,xj)。然后由于实际操作时，计算协方差阵要求$\Sigma\ \Phi(x) = 0$ ，所以我们得对核矩证进行修正：

K’(xi,xj)=

$$K(xi,xj)—\frac{1}{N}\sum_{m=1}^{N}\left(K(xi,xm)+K(xm,xj)\right)+\frac{1}{N*N}\sum_{m=1}^{N}\sum_{n=1}^{N}K(xm,xn)$$

然后分别求出 10 个核矩阵的最大的 p 个特征值及其对应的特征向量。

这样对于任意一个未知的图片样本，分别将其在这 10 组特征向量上投影。

$$Ym=\sum_{j=1}^{N}\boldsymbol{\alpha_j^m}*\left(k'(x,xj)\right)$$

然后对这 p 个特征向量上的投影取平方和，得到一个量度。$Zn=\sum_{j=1}^{p}Ym*Ym$,n=1,…10。 比较这 10 个量度间的大小就可以得到未知数据与 10 类样本的相似程度。

- **对再生核函数的研究**

这里分别选用了 3 种不同的核函数实现，旨在分析再生核函数的一些性质。选取了指数径向基核函数 K1(x,y)=$\frac{1}{2}e^{-||x-y||}$，经频域相乘再经傅立叶逆变换得其一次卷积再生核函数为

$$K2(x,y)=\frac{1}{4}(1+||x-y||)e^{-||x-y||},$$

同理得其二次卷积核函数为

$$K3(x,y)=\frac{1}{16}(||x-y||^2+3||x-y||+3)e^{-||x-y||}。$$

分别采用 3 个不同阶数的核函数对数据进行识别得到数据如下：

核函数	K1(x,y)	K2(x,y)	K3(x,y)
识别率	92%	92.6%	93.7%

分析得到，随着核函数阶数的提升，识别率有了少量的增加，但差别并不是很大。与某参考文献中关于“由高阶可导得函数空间中设计得核函数，其模式分类效果比从低阶可导得函数空间中设计出来的核函数好”的论述一致。但是由于时间精力有限，不能进行深入的系统研究。可以作为以后的探讨话题，将其作为选择高效核函数的一个研究方向。

- **SVM 与 KPCA 的比较**

支持向量积和核主成分研究方法都是模式识别中的热门研究对象，都有着高效的识别效率。本次实验中相同核函数在 2 种方法中达到的识别率几乎相同。可见方法并不是关键，核函数才是其选择研究的重心。但是由于 SVM 拥有自己的函数库，可能采用的大多为 matlab 所擅长的矩阵运算，所以识别速度相当快。1000 个数据用时约为 1 秒。而自己编写的 KPCA 程序由于涉及较多的循环操作，耗时较多 1000 个数据的识别用了 1 分多钟，两者的差距是明显的。排除程序优化带来的时间减小，可以估计 SVM 比 KPCA 具有较高的运算处理速率（基于 matlab 平台上）。

[观点]

1. 利用核函数将数据坐非线性运算映射到高维空间能够有效解决一些线性不可分的问题。
2. 不同的核函数的识别效果是不同的。
3. 同一种核函数的参数选择不同也会给识别带来一定的影响。
4. 训练样本的个数是影响 SVM 识别率的一大关键因素。
5. 目标特征提取方法的选择和识别的前期数据处理对识别效率也有着不小的影响。
6. 适当选择再生核函数的阶次有助于获得较好的识别率。

[总结]

为了做好这次 project，自己搜集查询了大量的参考资料，并前后多次与同学进行了交流探讨。前期主要是进行核函数的理论研究，试图弄清核函数的作用原理及其性质。中期进行了 matlab 软件的编程研究，对具体的数字识别课题进行研究，花了部分时间在文件数据的读取、分类、处理上，更多的时间则是消耗在 KPCA 的程序编写及 SVM 的正确使用上。最后则进入了平和的分析整理阶段。不断的调整自己的核函数和编程细节上，以达到较好的识别率。经过反复的试验比较，最后确立了具体的核函数方向，最终识别达到 0.941。

通过这次 project，我揭开了核方法这一热门研究的面纱。了解了其在模式识别中重要作用。并从 SVM 和 KPCA 两个方面分别体验了核方法的魅力，并取得了较高的识别率。

最后通过试验估计了获取高效核函数的一种可行的途径——选择适当阶次的再生核函数。

[致谢]

感谢无 41 的卢明同学提供了四字班的优秀 project 报告示例，使自己对什么样的报告才是一个好报告有了直观的认识。并学到四字班 3 位优秀同学的研究方法，并参考了其部分理论知识。

感谢耿泉同学在 SVM 方面给自己带来的小小启示。

感谢本班的张鑫同学在 KPCA 的研究过程中和自己进行的多次交流讨论。

[参考文献]

[1] B.Schölkopf, “Introduction to Kernel Methods”,

[2] J-P. Vert, et al “A primer on kernel methods”, Chapter 1 in “Kernel Methods in Computational Biology”, Edited by Bernhard Sch ölkopf, Koji Tsuda and Jean-Philippe Vert, MIT Press, 2004

[3] Shu Yang, et al “Bilinear Analysis for Kernel Selection and Nonlinear Feature Extraction” IEEE Trans on Neural Network, Vol 18, No.5, pp1442-1452, Sept, 2007

[4] Framework for Choice of Kernels W AN Hai—ping，HE H ua—can

[5] 核函数的性质及其构造方法 王国胜+基

[6] 基于核函数的学习算法 田盛丰

[7] SVM 在车牌字符识别中的应用 黄 凡，李志敏，张 晶，万 睿，张凤阳

[8] 改进的 RBF 网络及其参数优化方法 林成荫 高大启

[9] 从 RBF 核函数中抽取关联分类规则 邓正宏，张 阳，宋 群

[10] 模式分析的核函数设计方法及应用 柳桂国 ， 柳 贺 ， 黄道

[11] 于变异函数的径向基核函数参数估计 阎 辉 张学工 马云潜 李衍达