基于条件正定核的判决分类算法研究

王 源[1,2],陈亚军[3],蔡 彪[1]
(1.西华师范大学 微机应用研究所,四川 南充 637002;
2.淮南师范学院 信息技术系,安徽 淮南 232001;
3.西华师范大学 物理与电子信息学院,四川 南充 637002)

摘 要:在非线性分类算法中,重要的技术之一是对于核技巧的使用,目前仍是尚未解决的问题。针对 Mercer 核约束条件强的特点,引入条件正定核函数(c.p.d.核)改善其约束条件。实验证明,对于 KPCA 和 KDDA 等判决分类方法,通过 c.p.d.核的使用,仍能保持其固有分类性能,从而推广了此类方法的适用范围。

关键词:非线性分类算法;核主元分析;核直接判决分析;条件正定核
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2006)08-0051-04

A Study of Kernel Discriminant Classify Algorithms Based on CPD Kernels

WANG Yuan[1,2], CHEN Ya-jun[3], CAI Biao[1]
(1. Institute of Microcomputer Application, China West Normal University, Nanchong 637002, China;
2. Information Technology Department, Huainan teachers college, Huainan 232001, China;
3. School of Physics & Electronics, China West Normal University, Nanchong 637002, China)

Abstract: One of the important techniques is the trick of kernels among the nonlinear classify algorithms. It is a problem which cannot be solved completely till now. For the more rigid restriction conditions of Mercer kernels, substitute them with conditionally positive kernels. The latter has less conditions so as to develop its usage. Some experiments show that c.p.d kernels can be used for several classical nonlinear discriminant classify algorithms such as KPCA and KDDA. As a result, it extends the scope of these classify algorithms.

Key words: nonlinear classify algorithms; KPCA; KDDA; CPD kernel

0 引 言

随着统计学习理论在模式分类中的应用,核方法的使用成为非线性分类算法的一个重要方面。支持向量机的一个重要成分就是在高维特征空间中,利用输入数据的简单函数分级内积的计算来确定类属关系。在核方法基础上发展而来的各种非线性判决分类算法已经成为机器学习领域最有效的技术之一。

目前在实际使用中,主要还是以多项式核、径向基核(RDF)和两层神经网络的方式来使用核方法[1],过强的约束条件往往使得它们的使用范围受到极大的限制。文中通过对判决分析算法和核函数自身性质的分析,引入条件正定核函数来改善核方法的适用条件,从而拓宽其应用领域。

定义1:非线性变换 $\phi: \boldsymbol{X} \to \boldsymbol{F}, x_i \in \boldsymbol{X}, \phi(x_i) \in \boldsymbol{F}$ 是输入空间 $\boldsymbol{X}$ 到特征空间 $\boldsymbol{F}$ 的一个特征映射。

定义2:(核函数)在特征映射 ϕ 下的对称函数 $K: \boldsymbol{X} \times \boldsymbol{X} \to \mathbf{R}$,即

$$K(x_i, x_j) = \phi(x_i)\phi(x_j) \tag{1}$$

称为核函数。

条件1:(Mercer 条件)[2] 对于 $\mathbf{R}^N$ 上任意有界函数 $g(x)$,对称函数 $K(x,y)$ 满足

$$\int K(x,y)g(x)g(y)\mathrm{d}x\mathrm{d}y \geqslant 0$$

称该函数满足 Mercer 条件。

定义3:(正定核矩阵或 Mercer 核,positive definite kernel,p.d.)对于任意 $m \in \mathbf{N}, x_i \in \boldsymbol{X}$,上述核函数构成的 Gram 矩阵 $\boldsymbol{K}_{ij}$ 如果对于任意的 $c_i \in \mathbf{R}$ 均有

$$\sum_{i,j=1}^{m} c_i c_j \boldsymbol{K}_{ij} \geqslant 0 \tag{2}$$

即满足条件1,称为正定核。

1 KPCA(核主元分析)

对于训练样本集 $X = \{x_1, x_2, \cdots, x_n\}$,不失一般性,

收稿日期:2006-02-20
基金项目:四川省教育厅重点项目(2004A102)
作者简介:王 源(1971-),男,安徽淮南人,硕士,讲师,研究方向为机器学习。

假设 $E(x)=0$(如均值不为0,可用 $x=x-E(x)$),则特征子空间的基就是协方差矩阵

$$C=\frac{1}{n}\sum_{i=1}^{n}x_ix_i^{\mathrm{T}} \tag{3}$$

的特征向量。

对(3)式两端做特征映射 ϕ 得:

$$\bar{C}=\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi(x_i)^{\mathrm{T}}$$

$\bar{C}$ 的特征向量 V 就是原训练样本集的非线性特征子空间的基向量,即满足式:

$$\bar{C}V=\lambda V \tag{4}$$

特征向量 V 可表示为:

$$V=\sum_{i=1}^{n}\alpha_i\phi(x_i) \tag{5}$$

将(4)式两边乘 $\phi(x_i)$,得到:

$$\phi(x_i)\bar{C}V=\lambda\phi(x_i)V \tag{6}$$

将(5)式代入(6)式得:

$$K\alpha=n\lambda\alpha \tag{7}$$

$$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)^{\mathrm{T}}$$

解(7)式得 α,即由 $\bar{C}$ 的特征向量组成的矩阵 $V=(v_1,v_2,\cdots,v_n)$,也即非线性特征子空间。x_i 在非线性特征子空间上的子像,记为 y_i:

$$y_i=V^{\mathrm{T}}\phi(x_i)=\begin{pmatrix}\sum_{j=1}^{n}(\alpha_{1j}\phi(x_j)^{\mathrm{T}}\cdot\phi(x_i))\\ \sum_{j=1}^{n}(\alpha_{2j}\phi(x_j)^{\mathrm{T}}\cdot\phi(x_i))\\ \vdots\\ \sum_{j=1}^{n}(\alpha_{nj}\phi(x_j)^{\mathrm{T}}\cdot\phi(x_i))\end{pmatrix}=\begin{pmatrix}\sum_{j=1}^{n}(\alpha_{1j}K(x_j,x_i))\\ \sum_{j=1}^{n}(\alpha_{2j}K(x_j,x_i))\\ \vdots\\ \sum_{j=1}^{n}(\alpha_{nj}K(x_j,x_i))\end{pmatrix}$$

2 KDDA(核直接判别分析)

设 x_{ij}(n 维列向量)表示第 i 类目标的第j 个训练样本的一维像,利用上述的特征映射 ϕ 将输入空间X 上的数据样本非线性映射到特征空间 F 中。在 F 中定义准则函数为[3]:

$$J(w)=\frac{w^{\mathrm{T}}S_B^{\phi}w}{w^{\mathrm{T}}S_W^{\phi}w},w\in F$$

其中

$$w=\sum_{i=1}^{N}\alpha_i\phi(x_i) \tag{8}$$

经特征变换后在特征空间 F 中得到类间散度矩阵 S_B^{ϕ} 和类内散度矩阵 S_W^{ϕ} 分别为:

$$S_B^{\phi}=\sum_{i=1}^{g}(\overline{\phi(x_i)}-\overline{\phi(x)})(\overline{\phi(x_i)}-\overline{\phi(x)})^{\mathrm{T}}$$

$$S_W^{\phi}=\sum_{i=1}^{g}\sum_{j=1}^{N_i}(\phi(x_{ij})-\overline{\phi(x_i)})(\phi(x_{ij})-\overline{\phi(x_i)})^{\mathrm{T}}$$

令 $\overline{\phi(x_i)}=\frac{1}{N_i}\sum_{j=1}^{N_i}\phi(x_{ij})$ (9)

及 $\overline{\phi(x)}=\frac{1}{N}\sum_{i=1}^{R}\sum_{j=1}^{N_i}\phi(x_{ij})$ (10)

其中 $i=1,2,\cdots,g;j=1,2,\cdots,N_i;N=N_1+N_2+\cdots+N_g$;$g$ 为目标类别数,N_i 为第i 类目标的训练样本数,N 为训练样本总数。

显然 $J(w)$ 越大,表明类间差异相对于类内差异越大,越有利于目标分类。

将(9)式和(10)式两端分别乘(8)式的转置得到:

$$w^{\mathrm{T}}\overline{\phi(x_i)}=\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N_i}\alpha_jK(x_j,x_{ik})=\alpha^{\mathrm{T}}P_i$$

$$w^{\mathrm{T}}\overline{\phi(x_i)}=\frac{1}{NN_i}\sum_{i=1}^{g}\sum_{j=1}^{N}\sum_{k=1}^{N_i}\alpha_jK(x_j,x_{ik})=\alpha^{\mathrm{T}}Q$$

$$w^{\mathrm{T}}\phi(x_{ij})=\sum_{k=1}^{N}\alpha_kK(x_k,x_{ij})=\alpha^{\mathrm{T}}(K_k)_{ij}$$

而$(P_i)_j=\frac{1}{N_i}\sum_{k=1}^{N_i}K(x_j,x_{ik})$,

$$Q_j=\frac{1}{NN_i}\sum_{i=1}^{g}\sum_{k=1}^{N_i}\alpha_jK(x_j,x_{ik})$$

令 $w^{\mathrm{T}}S_B^{\phi}w=\alpha^{\mathrm{T}}R_B\alpha,w^{\mathrm{T}}S_W^{\phi}w=\alpha^{\mathrm{T}}R_W\alpha$

则 $R_B=\sum_{i=1}^{K}(P_i-Q)(P_i-Q)^{\mathrm{T}}$

$$R_W=\sum_{i=1}^{g}\sum_{j=1}^{N_i}((K_k)_{ij}-P_i)((K_k)_{ij})-P_i)^{\mathrm{T}}$$

这样,最大化原准则函数 $J(w)$ 可等价为最大化 $J(\alpha)=\frac{\alpha^{\mathrm{T}}R_B\alpha}{\alpha^{\mathrm{T}}R_W\alpha}$。

可求得矩阵 $R_W^{-1}R_B$ 的一组基 α,由此可得 w,构成非线性变换矩阵 V,即为非线性判别子空间。x_i 在非线性判别子空间的子像为:

$$y_i=V^{\mathrm{T}}\phi(x_i)=\begin{pmatrix}\sum_{j=1}^{n}(\alpha_{1j}\phi(x_j)^{\mathrm{T}}\cdot\phi(x_i))\\ \sum_{j=1}^{n}(\alpha_{2j}\phi(x_j)^{\mathrm{T}}\cdot\phi(x_i))\\ \vdots\\ \sum_{j=1}^{n}(\alpha_{nj}\phi(x_j)^{\mathrm{T}}\cdot\phi(x_i))\end{pmatrix}=\begin{pmatrix}\sum_{j=1}^{n}(\alpha_{1j}K(x_j,x_i))\\ \sum_{j=1}^{n}(\alpha_{2j}K(x_j,x_i))\\ \vdots\\ \sum_{j=1}^{n}(\alpha_{nj}K(x_j,x_i))\end{pmatrix}$$

因而在核主元分析和核直接判别分析中,不论是求解 α_i,还是求解子像 y_i,只需在原始训练样本空间中计算函

数 K 的各分级内积,而无需知道 $\phi(x)$ 的具体表达式。对非线性分类算法的构造也就转化为对核函数自身性质的研究。

3　条件正定核的使用

在现实中,寻找合适的满足条件 1 的核一般很难,考查满足 Mercer 条件的核函数或等价的正定核: $K(x_i,x_j)=\phi(x_i)\phi(x_j)$ 的形式,即

$$K(\boldsymbol{x},\boldsymbol{x}^{\mathrm{T}})=(\phi(\boldsymbol{x}),\phi(\boldsymbol{x}^{\mathrm{T}}))$$

说明核函数可以被看成一个简化的相似性测度(内积): $(\boldsymbol{y},\boldsymbol{y}^{\mathrm{T}}),\boldsymbol{y}\in\boldsymbol{F}$;而另一测度即距离(相异性测度) $\|\boldsymbol{y}-\boldsymbol{y}^{\mathrm{T}}\|^2$ 可表示为:

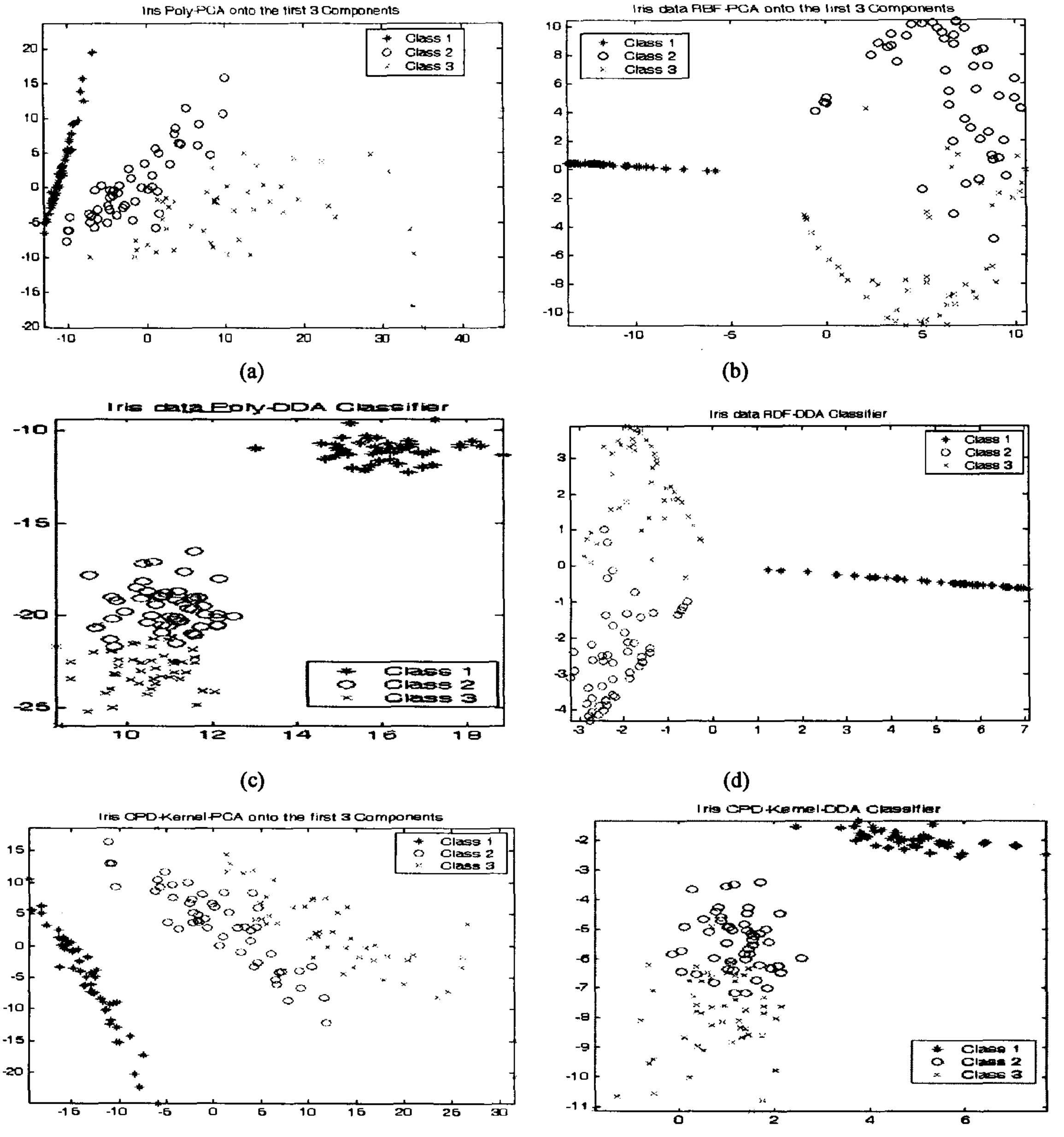
$$\|\phi(\boldsymbol{x})-\phi(\boldsymbol{x}^{\mathrm{T}})\|^2=(\phi(\boldsymbol{x})-\phi(\boldsymbol{x}^{\mathrm{T}}),\phi(\boldsymbol{x})-\phi(\boldsymbol{x}^{\mathrm{T}}))=K(\boldsymbol{x},\boldsymbol{x})+K(\boldsymbol{x}^{\mathrm{T}},\boldsymbol{x}^{\mathrm{T}})-2K(\boldsymbol{x},\boldsymbol{x}^{\mathrm{T}}) \tag{11}$$

即在特征空间 $\boldsymbol{F}$ 中距离测度与 Mercer 核之间存在联系。

定义 4:(条件正定核函数,Conditionally positive definite kernel,c. p. d.),若对称函数 $K^*:\boldsymbol{X}\times\boldsymbol{X}\rightarrow\mathbf{R}$ 对于所有的 $m\in\mathbf{N},x_i\in\boldsymbol{X}$,当 $\sum_{i=1}^{m}c_i=0$ 时满足(2)式,则称 K^* 为一个条件正定核(c. p. d. 核)。

定理 1　令 $x_0\in\boldsymbol{X},K^*$ 是 $\boldsymbol{X}\times\boldsymbol{X}$ 上的对称函数,则

$$K(\boldsymbol{x},\boldsymbol{x}^{\mathrm{T}})=K^*(\boldsymbol{x},\boldsymbol{x}^{\mathrm{T}})-K^*(\boldsymbol{x},\boldsymbol{x}_0)-K^*(\boldsymbol{x}_0,\boldsymbol{x}^{\mathrm{T}})+K^*(\boldsymbol{x}_0,\boldsymbol{x}_0) \tag{12}$$



(a)(c) 多项式核的 KPCA 和 KDDA 分类结果(d=0.5);　(b)(d) RDF 核的 KPCA 和 KDDA 分类结果($\sigma^2=0.5$);
(e)　c.p.d 核的 KPCA 分类结果(d=0.5);　(f)　c.p.d 核的 KDDA 分类结果 $\beta=1$。

图 1　各算法分类实验结果

是正定的当且仅当 K^* 是条件正定的。

将(12)式带入(11)式:存在c.p.d核 K^* 使得:

$$\|\phi(x)-\phi(x^T)\|^2=\frac{1}{2}(K^*(x,x)+K^*(x^T,x^T)-K^*(x,x^T)\qquad(13)$$

特别的,如果 $K^*(x,x)=0$(如在Hilbert空间表示下),(13)式可简化为:

$$\|\phi(x)-\phi(x^T)\|^2=-K^*(x,x^T)\qquad(14)$$

同时另一方面,有

$$(\phi(x)-\phi(x_0),\phi(x^T)-\phi(x_0))=\frac{1}{2}[-\|\phi(x)-\phi(x^T)\|^2+\|\phi(x)-\phi(x_0)\|^2+\|\phi(x_0)-\phi(x^T)\|^2]\qquad(15)$$

因而可以通过使用c.p.d核来代替直接计算核函数的分级内积,这样做的另一个优势是:由于内积运算没有变换不变性,而 $\|\cdot\|^2$ 运算具有平移等变换不变性,因而可以具有更大的适用范围。

事实上[5],任何负的平方距离核函数均是条件正定而非正定的,即如下形式的核[4]均是c.p.d.核(而非p.d.核):

$$K^*(x,x^T)=-\|x-x^T\|^{\beta}\qquad(16)$$

Smola等人证明[5]只有 $0<\beta<2$ 的(16)式才适用于构造非线性分类算法。

4 实验及分析

为了验证文中所述算法的有效性,笔者使用标准测试集:鸢尾属植物数据集(iris.dat)进行了分类实验。在实验中分别使用多项式核函数和径向基核函数对数据集的4种属性(萼片长度、萼片宽度、花瓣长度、花瓣宽度)的3个类别(setosa,versicolor,virginica)进行KPCA和KDDA分析,引入 $\beta=1$ 的条件正定核后,利用该类型核的组合执行上述核分类算法,并对其分类性能进行对比测试,结果如图1所示。实验结果表明使用该c.p.d核与多项式核具有等价的效果。由于采用的c.p.d核是线性距离核,其整体的分类效果不如使用径向基核。

由于KPCA和KDDA在图像处理、语音和信号处理、模式识别、机器学习等领域有着广泛的用途,上述实验也表明使用c.p.d核同样能够完成这类非线性分类和识别的任务。

5 结束语

随着模式识别与机器学习技术的发展,及越来越多复杂分类任务的提出,使得简单线性分类算法或其组合均难以胜任这种要求。核策略等非线性方法的使用有效地扩展了学习算法。通过对条件正定核的使用,说明了基于核距离的非线性算法较传统距离度量有更强的适应性和有效性,比Mercer核有着更宽松的约束条件,因而有效地扩展了核方法的适用领域。

参考文献:

[1] 孙即祥.现代模式识别[M].长沙:国防科技大学出版社,2002.

[2] Duda R O,Hart P E,Stork D G.模式分类(第2版)[M].李宏东,姚天翔译.北京:机械工业出版社,2003.

[3] Lu Juwei. Face Recognition Using Kernel Direct Discriminant Analysis Algorithms[J]. IEEE TRANSACTIONS ON NEURAL NETWORKS,2003,14(1):117-126.

[4] 孔 锐,张国宣.基于核的K-均值聚类[J].计算机工程,2004(11):12-13.

[5] Smola A J,Schölkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion [R]. GMD Technical Report,GMD- FIRST,Berlin:[s.n.],1998.1064-1097.

(上接第50页)

```
        break;
      }
   }
   //释放线程池内存
   delete pThreadPool;
}
```

3 结束语

线程池是组织服务器应用程序的有用工具,可以显著改善服务器程序的性能,在服务器领域有着广泛的应用前景。文中在分析了传统的多线程服务器软件处理轻量级任务的弊端的基础上,提出了一个基于线程池技术的服务器软件模型;同时给出了一个通用的可扩展线程池框架,以及其在服务器软件中应用的具体实现。基于线程池技术的服务器软件模型,对于提高需要处理轻量级任务并且实时性要求较高的服务器性能有很大帮助,具有现实意义。

参考文献:

[1] Beveridge J,Wiener R. Win32多线程程序设计[M].侯 捷译.武汉:华中科技大学出版社,2002.

[2] 王险峰,刘宝宏. Windows环境下的多线程编程原理与应用[M].北京:清华大学出版社,2002.

[3] Butenhof D R. POSIX多线程程序设计[M].于 磊,曾刚译.北京:中国电力出版社,2003.

[4] Hughes C,Hughes T. C++面向对象多线程编程[M].周良忠译.北京:人民邮电出版社,2001.

[5] Richter J. Windows高级编程指南[M].王书洪,刘光明译.北京:清华大学出版社,1999.