

最优双核复合分类算法的构造

王 峰, 张鸿宾

(北京工业大学计算机学院, 北京 100124)

摘 要: 由于使用单一且固定的核函数, 传统的核分类算法不能有效地适应复杂的数据集合, 导致分类性能下降. 本文提出一种基于双核复合的分类算法 ODKC (Optimal Double-Kernel Combination) 的构造框架, 通过融合两个基本核函数的映射来构造目标核函数. 研究了双核复合的三种典型方式, 并把这三种复合方式纳入到统一的框架下处理. 论文以核与数据的匹配性度量 KTA (Kernel Target Alignment) 以及分类性能验证了所提算法的有效性.

关键词: 核方法; 双核复合; 分类

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2012)02-0260-06

电子学报 URL: <http://www.ejournal.org.cn> **DOI:** 10.3969/j.issn.0372-2112.2012.02.009

Optimal Double-Kernel Combination Method for Classification

WANG Feng, ZHANG Hong-Bin

(College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: Traditional kernelised classification methods could not perform well sometimes because of using a single and fixed kernel, especially on some complicated data sets. In this paper, a novel optimal double-kernel combination (ODKC) method is proposed for complicated classification tasks. Firstly, data are mapped by two basic kernels into different feature spaces respectively, and then three kinds of optimal composite kernels are constructed by integrating information of the two feature spaces. Comparative experiments demonstrate the effectiveness of our methods.

Key words: kernel methods; double-kernel combination; classification

1 引言

近年来核方法在模式识别等领域得到了广泛的应用, 核方法与经典线性算法相结合, 产生了一系列的核化算法, 如支撑向量机 (SVM)^[1], 核 Fisher 判别 (KFD)^[2], 核判别分析 (KDA)^[3], 核主成分分析 (KPCA)^[4], 核典型相关分析 (KCCA)^[5], 核独立分量分析 (KICA)^[6], 核最小均方误差 (KMSE)^[7] 等. 这些方法及相应改进算法在使用时^[8,9], 首先要选择一个合适的核函数, 将低维数据映射到高维的特征空间, 然后在特征空间中进行线性分类或特征选择等运算. 这些算法取得好的效果的关键在于所选择的核函数适合目标数据, 使映射之后数据集的线性可分性得到改善. 然而, 由于使用了单一且固定的基本核函数, 传统的核化分类算法对一些复杂任务的适应能力仍然有待提高^[10-12].

为克服使用单一核的局限性, 近年来出现了面向特定任务、利用多个基本核构造新的目标核函数的核优化方法. Lanckriet 等利用多个核函数的凸组合来构造目标

核函数, 通过求解二次规划问题优化目标核^[10]. Bach 针对文献[10]对大规模数据处理上的局限^[11], 提出了文献[10]中二次规划问题的对偶形式. Bennett^[12]和 Bi^[13]分别提出了 Boost 算法用于实现多种基本核的复合. Xiong 等使用了数据依赖核 (Data-Dependent kernel) 作为目标核函数^[14], 通过最大类间分离性度量来构造和优化目标核. Chen 等分析了文献[14]中使用的准则函数的局限性, 提出了一种统一的核优化方法的框架^[15]. 然而, 文献[14, 15]的方法都需要更多的标签样本, 一部分用来构造数据依赖核, 另一部分则用来训练. 为了避免特征值求解, 文献[14, 15]都使用了迭代算法, 迭代的次数需要事先给出. 另外, Cristianini 等提出了衡量核函数“好坏”的度量, 即核与数据间的匹配性度量 (Kernel Target Alignment, KTA)^[16]. KTA 值越高, 说明核函数与目标数据越适合, 并以此来构造新的核函数. 除了基本核的组合之外, 近年来也出现了一些组合多尺度核以及无限核的研究^[17, 18].

与传统的核方法相比, 上述方法在对复杂数据的

适应性上有了一定的改善.但从计算结构上分析,它们仍属于浅层计算结构,仍然存在浅层结构学习的局限性,对复杂多变的分界面函数的学习能力有限^[19].本文针对实际工作中经常使用的两个核的复合问题,提出一种最优双核复合分类方法(ODKC)来构造目标核函数,分析了三种典型的核复合方式,并将它们纳入到统一框架下进行讨论.利用核与数据的匹配性度量 KTA 和分类性能实验,比较了 ODKC 和单一核的性能.

2 基本思想

给定一组有标签的训练数据集 $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in \mathbf{R}^n$, $y_i \in \{1, \dots, m\}$ 是类别标签, m 为类别个数, N 为样本总数.目标是要构造一个最优的双核复合的核映射 $\Phi(\mathbf{x}) = \phi_1(\phi_0(\mathbf{x}))$ (ODKC 特征映射),使得图 1(a)中原本线性不可分的输入数据在特征空间中能够变得类内紧致和类间分离(图 1(c)).ODKC 特征映射由两个非线性特征映射 ϕ_0 和 ϕ_1 复合而成.其中 ϕ_0 是一个基本的核映射.对于一些复杂的数据集,数据经过 ϕ_0 映射后,仍然可能线性不可分,如图 1(b)所示.因此,有必要再进行一次特征映射,重构数据的几何结构,使之能更好地分类.我们选择 $\mathbf{Q}^T \phi_2(\cdot)$ 作为 ϕ_1 ,其中 ϕ_2 是一个非线性映射, \mathbf{Q}^T 为线性算子. ϕ_2 的构造决定了两个非线性特征映射的复合形式.本文将讨论 ϕ_2 的三种构造形式以及如何根据准则函数优化算子 \mathbf{Q}^T .下面先引入核优化的准则函数.

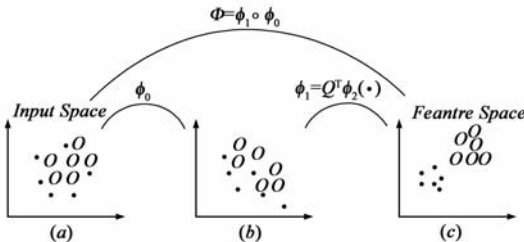


图1 输入空间(a)中的两类数据“•”和“○”经过 $\Phi(\mathbf{x})$ 映射到特征空间(c)

准则函数 ODKC 利用训练集类别信息,以增强特征空间中数据的类内紧致性和类间分离性为目标,通过优化如下准则函数求解两个非线性特征映射的最优复合.先引入样本的类别隶属矩阵:

定义 1 称矩阵 $\mathbf{A} = [\delta_{ij}]_{m \times N}$ 为类别隶属矩阵,其元素为:

$$\delta_{ij} = \begin{cases} 0, & f(\mathbf{x}_j) = i \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

其中 $f(\mathbf{x})$ 为标签函数,当且仅当 \mathbf{x}_j 的标签为 i 时, $f(\mathbf{x}_j) = i, 1 \leq i \leq m$, 矩阵 \mathbf{A} 包含了训练样本的类别信息.

本文采用如下的准则函数来刻画样本的类内紧致性和类间分离性:

$$\begin{aligned} \min_{r_i, \Phi} g(r_i, \Phi) &= \sum_{i=1}^m r_i^2 + \sum_{i=1}^m \sum_{j=1}^N \lambda_{ij} (\|\mathbf{c}_i^\Phi - \Phi(\mathbf{x}_j)\|^2 - r_i^2) \delta_{ij} \\ &= \sum_{i=1}^m r_i^2 + \sum_{i=1}^m \sum_{j=1}^N \|\mathbf{c}_i^\Phi - \Phi(\mathbf{x}_j)\|^2 \lambda_{ij} \delta_{ij} \\ &\quad - \sum_{i=1}^m \sum_{j=1}^N r_i^2 \lambda_{ij} \delta_{ij} \end{aligned} \quad (2)$$

其中, $\mathbf{c}_i^\Phi = \frac{1}{N_i} \sum_{f(\mathbf{x}_j)=i} \Phi(\mathbf{x}_j)$ 是在特征映射 $\Phi(\cdot)$ 对应的特征空间中第 i 类数据的类中心, N_i 表示第 i 类数据的个数; $r_i^2 = \max_{\mathbf{x}_j} \{\|\Phi(\mathbf{x}_j) - \mathbf{c}_i^\Phi\|^2 | f(\mathbf{x}_j) = i\}$ 是特征空间中第 i 类的数据 $\Phi(\mathbf{x}_j)$ 到第 i 类中心的最大距离的平方.准则函数中的权矩阵 $\mathbf{A} = (\lambda_{ij})_{m \times N}$ 表示当数据点靠近异类数据中心时的惩罚因子,它的选取将在 3.4 节讨论.

显然,式(2)最小化问题的解 $(r_i^2, \mathbf{c}_i^\Phi, \Phi)$ 使得特征空间中包含每一类数据的超球面具有最小的半径,即类内样本更紧致,而不同类的样本相互远离.

3 复合核映射 Φ 及其核矩阵 \mathbf{K}^Φ 的求解

对于式(2)的准则函数,下面讨论复合目标核映射 Φ 及其核矩阵 \mathbf{K}^Φ 的求解.

3.1 最小化准则函数

在最小化这个准则函数之前,先对式(2)进行化简.令 $\mathbf{c}_i^{\phi_0} = \frac{1}{N_i} \sum_{f(\mathbf{x}_j)=i} \phi_0(\mathbf{x}_j)$, 矩阵 $\mathbf{F} = (\lambda_{ij} \delta_{ij})_{m \times N}$, 则式(2)右边第二项可化为:

$$\sum_{i=1}^m \sum_{j=1}^N \text{trace} \left[\left(\mathbf{Q}^T \phi_2(\mathbf{c}_i^{\phi_0}) - \mathbf{Q}^T \phi_2(\phi_0(\mathbf{x}_j)) \right) \cdot \left(\mathbf{Q}^T \phi_2(\mathbf{c}_i^{\phi_0}) - \mathbf{Q}^T \phi_2(\phi_0(\mathbf{x}_j)) \right)^T \right] F_{ij}$$

若记 $\mathbf{C} = (\phi_2(\mathbf{c}_1^{\phi_0}), \phi_2(\mathbf{c}_2^{\phi_0}), \dots, \phi_2(\mathbf{c}_m^{\phi_0}))$
 $\phi_2(\phi_0(\mathbf{X})) = (\phi_2(\phi_0(\mathbf{x}_1)), \phi_2(\phi_0(\mathbf{x}_2)), \dots, \phi_2(\phi_0(\mathbf{x}_N)))$
 $\mathbf{A} = \mathbf{C} \mathbf{S} \mathbf{C}^T - 2 \mathbf{C} \mathbf{F} \phi_2^T(\phi_0(\mathbf{X})) + \phi_2(\phi_0(\mathbf{X})) \bar{\mathbf{S}} \phi_2^T(\phi_0(\mathbf{X}))$
 其中, $\bar{\mathbf{S}}_{m \times m}$ 和 $\hat{\mathbf{S}}_{N \times N}$ 是对角阵, $\bar{\mathbf{S}}_{m \times m}$ 对角线上的元素为矩阵 \mathbf{F} 的该行元素之和, $\hat{\mathbf{S}}_{N \times N}$ 对角线上的元素为矩阵 \mathbf{F} 的该列元素之和,则最终可得:

$$\sum_{j=1}^s \mathbf{q}_j^T \mathbf{A} \mathbf{q}_j \quad (3)$$

其中, \mathbf{q}_j 是矩阵 \mathbf{Q} 的第 j 列.将式(3)代入式(2),考虑约束 $\mathbf{q}_j^T \mathbf{q}_j = 1$, 得如下化简的最优化问题:

$$\begin{aligned} \min_{r_i, \mathbf{q}_j} \sum_{i=1}^m r_i^2 + \sum_{j=1}^s \mathbf{q}_j^T \mathbf{A} \mathbf{q}_j - \sum_{i=1}^m \sum_{j=1}^N r_i^2 \lambda_{ij} \delta_{ij} \\ \text{s.t.} \quad \mathbf{q}_j^T \mathbf{q}_j = 1 \end{aligned}$$

定义下面的拉格朗日函数求解这个最优化问题,即

$$L(r_i, \mathbf{q}_j, \beta_j) = \sum_{i=1}^m r_i^2 + \sum_{j=1}^s \mathbf{q}_j^T \mathbf{A} \mathbf{q}_j - \sum_{i=1}^m \sum_{j=1}^N r_i^2 \lambda_{ij} \delta_{ij}$$

$$- \sum_{j=1}^s \beta_j (\mathbf{q}_j^T \mathbf{q}_j - 1) \quad (4)$$

$\beta_j (j=1, \dots, s)$ 为拉格朗日乘子. 将式(4)分别对 $\mathbf{q}_j (j=1, \dots, s)$ 和 $r_i (i=1, \dots, m)$ 求偏导并令其为零得:

$$\frac{\partial \mathbf{L}}{\partial \mathbf{q}_j} = 2(\mathbf{A}\mathbf{q}_j - \beta_j \mathbf{q}_j) = 0 \quad (5)$$

$$\frac{\partial \mathbf{L}}{\partial r_i} = r_i \left(1 - \sum_{j=1}^N \lambda_{ij} \delta_{ij} \right) = 0 \quad (6)$$

由式(5)可得 $\mathbf{A}\mathbf{q}_j = \beta_j \mathbf{q}_j$, 可以看出拉格朗日乘子 β_j 正是矩阵 \mathbf{A} 的特征值, 而 \mathbf{q}_j 正是 \mathbf{A} 的特征向量. 将式(5), (6)代入式(4), 可以得到 $\mathbf{L}(r_i, \mathbf{q}_i, \beta_i) = \sum_{i=1}^s \beta_i$. 为了最小化该式, 设 s 为矩阵 \mathbf{A} 的负特征值的个数, 选择这些负特征值及其对应的特征向量作为最优解.

因此, 最优目标核映射 Φ 为

$$\Phi(\cdot) = \mathbf{Q}^T \phi_2(\phi_0(\cdot)) = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_s] \phi_2(\phi_0(\cdot))$$

其中 $\mathbf{q}_j (j=1, \dots, s)$ 为矩阵 \mathbf{A} 的负特征值对应的特征向量. 然而, 从矩阵 \mathbf{A} 的构造过程可以看出, 由于特征映射 ϕ_0 和 ϕ_2 是隐式的, 所以构造矩阵 \mathbf{A} 从而计算 \mathbf{q}_j 是不可能的. 事实上, 为求最优目标核映射 Φ 及其核矩阵 \mathbf{K}^Φ , 构造矩阵 \mathbf{A} 也是没有必要的. 下面我们讨论目标核映射 Φ 及其核矩阵 \mathbf{K}^Φ 的实际求解.

3.2 目标核映射 Φ 及其核矩阵 \mathbf{K}^Φ 的求解

定义向量 $\mathbf{v}_i \in \mathbf{R}^N$, 其第 k 个元素为:

$$(v_i)_k = \begin{cases} \frac{1}{N_i}, & f(\mathbf{x}_k) = i \\ 0, & \text{otherwise} \end{cases}$$

记矩阵 $\mathbf{V}_{N \times m} = (v_1, v_2, \dots, v_m)$, 可得

$$\begin{aligned} \mathbf{C} &= \phi_2(\phi_0(\mathbf{X})) \mathbf{V}, \\ \mathbf{A} &= \phi_2(\phi_0(\mathbf{X})) (\mathbf{V} \hat{\mathbf{S}} \mathbf{V}^T - \mathbf{V} \mathbf{F} - \mathbf{F}^T \mathbf{V}^T + \hat{\mathbf{S}}) \\ &\quad \cdot \phi_2^T(\phi_0(\mathbf{X})) \end{aligned} \quad (7)$$

$$\text{将式(7)代入式(5), 有: } \phi_2(\phi_0(\mathbf{X})) (\mathbf{V} \hat{\mathbf{S}} \mathbf{V}^T - \mathbf{V} \mathbf{F} - \mathbf{F}^T \mathbf{V}^T + \hat{\mathbf{S}}) \cdot \phi_2^T(\phi_0(\mathbf{X})) \mathbf{q}_j = \beta_j \mathbf{q}_j \quad (8)$$

$$\text{从而 } \mathbf{q}_j = \phi_2(\phi_0(\mathbf{X})) (\mathbf{V} \hat{\mathbf{S}} \mathbf{V}^T - \mathbf{V} \mathbf{F} - \mathbf{F}^T \mathbf{V}^T + \hat{\mathbf{S}}) \cdot \phi_2^T(\phi_0(\mathbf{X})) \mathbf{q}_j \frac{1}{\beta_j}$$

$$\text{令 } \boldsymbol{\alpha}_j = (\mathbf{V} \hat{\mathbf{S}} \mathbf{V}^T - \mathbf{V} \mathbf{F} - \mathbf{F}^T \mathbf{V}^T + \hat{\mathbf{S}}) \phi_2^T(\phi_0(\mathbf{X})) \mathbf{q}_j \frac{1}{\beta_j} \quad (9)$$

则有 $\mathbf{q}_j = \phi_2(\phi_0(\mathbf{X})) \boldsymbol{\alpha}_j$
将式(9)代入式(8), 并在式(8)两边左乘 $\phi_2^T(\phi_0(\mathbf{X}))$, 可得

$$\mathbf{K}_{\phi_2 \circ \phi_0} (\mathbf{V} \hat{\mathbf{S}} \mathbf{V}^T - \mathbf{V} \mathbf{F} - \mathbf{F}^T \mathbf{V}^T + \hat{\mathbf{S}}) \mathbf{K}_{\phi_2 \circ \phi_0} \boldsymbol{\alpha}_j = \beta_j \mathbf{K}_{\phi_2 \circ \phi_0} \boldsymbol{\alpha}_j \quad (10)$$

其中, $\mathbf{K}_{\phi_2 \circ \phi_0} = \phi_2^T(\phi_0(\mathbf{X})) \phi_2(\phi_0(\mathbf{X}))$ 是对应于特征映射 $\phi_2(\phi_0(\mathbf{X}))$ 的 Gram 矩阵. 因此, 给定一组输入数据 $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, 经过最优复合核特征映射 Φ 之后的数据应为:

$$\begin{aligned} &(\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}_N)) \\ &= (\mathbf{Q}^T \phi_2(\phi_0(\mathbf{x}_1)), \dots, \mathbf{Q}^T \phi_2(\phi_0(\mathbf{x}_N))) \\ &= (\boldsymbol{\alpha}_1^T \phi_2^T(\phi_0(\mathbf{X})), \dots, \boldsymbol{\alpha}_s^T \phi_2^T(\phi_0(\mathbf{X}))) \phi_2(\phi_0(\mathbf{X})) \\ &= (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_s)^T \mathbf{K}_{\phi_2 \circ \phi_0} \end{aligned} \quad (11)$$

这时最优化问题转变成了求解等式(10)的广义特征向量问题. 对于测试数据 \mathbf{x}_{test} , 可得:

$$\begin{aligned} &\Phi(\mathbf{x}_{test}) \\ &= \mathbf{Q}^T \phi_2^T(\phi_0(\mathbf{x}_{test})) \\ &= (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_s)^T (k_{\phi_2 \circ \phi_0}(\mathbf{x}_1, \mathbf{x}_{test}), \dots, k_{\phi_2 \circ \phi_0}(\mathbf{x}_N, \mathbf{x}_{test})) \end{aligned} \quad (12)$$

最优映射 $\Phi(\cdot)$ 对应的 Gram 矩阵为:

$$\begin{aligned} \mathbf{K}^\Phi &= (\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N))^T ((\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N))) \\ &= \mathbf{K}_{\phi_2 \circ \phi_0} (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_s) (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_s)^T \mathbf{K}_{\phi_2 \circ \phi_0} \end{aligned} \quad (13)$$

核矩阵 \mathbf{K}^Φ 在本文中也称为 ODKC 核矩阵. 由式(13)可以看出, 为求得核矩阵 \mathbf{K}^Φ , 须首先求得核矩阵 $\mathbf{K}_{\phi_2 \circ \phi_0}$. 下节我们讨论特征映射 ϕ_2 的构造和 $\mathbf{K}_{\phi_2 \circ \phi_0}$ 的计算.

3.3 特征映射 ϕ_2 的构造和 $\mathbf{K}_{\phi_2 \circ \phi_0}$ 的计算

本文利用双核复合的方法来构造目标核函数. 根据特征映射 ϕ_2 不同的定义, 讨论三种构造目标核函数的核函数复合方法, 并将这三种复合方法在一个统一的框架下进行讨论. 本文采用如下三种方式定义特征映射 ϕ_2 .

设 \mathbf{S} 为一有限样本集 $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$, ϕ_3 选取为不同于 ϕ_0 的另一个基本核特征映射, k_{ϕ_0} 和 k_{ϕ_3} 分别是特征映射 ϕ_0 和 ϕ_3 的核函数, N_1, N_2 分别是相应特征空间的维数, \mathbf{K}_{ϕ_0} 和 \mathbf{K}_{ϕ_3} 是相应的核矩阵, $k_{\phi_2 \circ \phi_0}$ 是对应于特征映射 $\phi_2(\phi_0(\mathbf{X}))$ 的核函数, 则三种复合方式的定义分别为:

定义 2 并联复合方式 特征映射 ϕ_2 定义为, $\phi_2(\phi_0(\mathbf{x})) \mapsto [\phi_0(\mathbf{x}), \phi_3(\mathbf{x})]$.

符号 $[\mathbf{v}_1, \mathbf{v}_2] = (v_{11}, \dots, v_{1N_1}, v_{21}, \dots, v_{2N_2})^T$ 表示向量 \mathbf{v}_1 和向量 \mathbf{v}_2 的连接. 在这种定义下, 核函数 $k_{\phi_2 \circ \phi_0}(x, z)$ 为: $k_{\phi_2 \circ \phi_0}(\mathbf{x}, \mathbf{z}) = \langle \phi_2(\phi_0(\mathbf{x})), \phi_2(\phi_0(\mathbf{z})) \rangle = k_{\phi_0}(\mathbf{x}, \mathbf{z}) + k_{\phi_3}(\mathbf{x}, \mathbf{z})$

所以, 核矩阵 $\mathbf{K}_{\phi_2 \circ \phi_0} = \mathbf{K}_{\phi_0} + \mathbf{K}_{\phi_3}$.

在这种定义的 ϕ_2 下, ODKC 核对应的等价计算结构如图 2 所示. 输入数据 \mathbf{x}_{new} 被 ϕ_0 和 ϕ_3 分别映射到两个特征空间中, 然后根据准则函数在两个特征空间中进行特征提取, 即优化权矩阵 \mathbf{Q} , 从而形成目标特征空间(图 2 第三层).

定义 3 乘法复合方式 特征映射 ϕ_2 定义为

$$\phi_2(\phi_0(\mathbf{x})) \mapsto (\phi_0(\mathbf{x}))_i (\phi_3(\mathbf{x}))_j$$

其中, $i=1, \dots, N_1, j=1, \dots, N_2, N_1$ 和 N_2 分别是 $\phi_0(\cdot)$

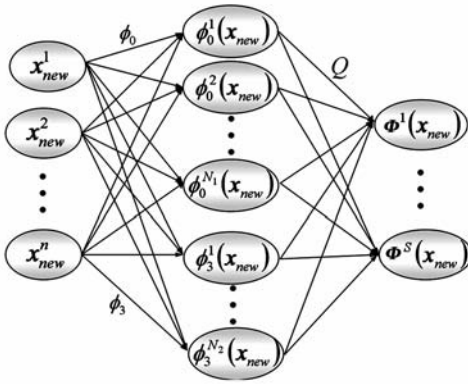


图2 在定义2的 ϕ_2 方式下,输入数据 \mathbf{x}_{new} 的复合映射 $\Phi(\cdot)$ 和 $\phi_3(\cdot)$ 的维数.则核函数 $k_{\phi_2 \circ \phi_0}(x, z)$ 为:

$$k_{\phi_2 \circ \phi_0}(x, z) = \langle \phi_2(\phi_0(x)), \phi_2(\phi_0(z)) \rangle = k_{\phi_0}(x, z) k_{\phi_3}(x, z)$$

所以,核矩阵 $(K_{\phi_2 \circ \phi_0})_{ij} = (K_{\phi_0})_{ij} * (K_{\phi_3})_{ij}$

在这种定义的 ϕ_2 下,ODKC核对应的等价计算结构如图3所示.输入数据 \mathbf{x}_{new} 被 ϕ_0 和 ϕ_3 映射到了两个特征空间中(图3第二层),然后融合不同特征空间属性间的相关性信息后构成另一更高维的特征空间(图3第三层),最后根据准则函数进行特征提取,从而形成目标特征空间(图2第四层).

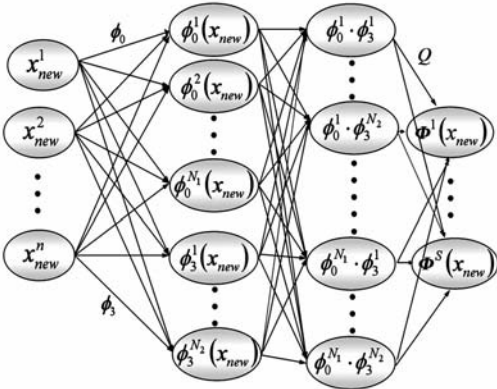


图3 在定义3的 ϕ_2 下,输入数据 \mathbf{x}_{new} 的复合特征映射 $\Phi(\cdot)$

定义4 串联复合方式 特征映射 ϕ_2 定义为 $\phi_2(\phi_0(x)) \mapsto \phi_3(\phi_0(x))$ 此时核函数 $k_{\phi_2 \circ \phi_0}(x, z)$ 为:

$$k_{\phi_2 \circ \phi_0}(x, z) = \langle \phi_2(\phi_0(x)), \phi_2(\phi_0(z)) \rangle = k_{\phi_3}(k_{\phi_0}(x, z))$$

所以,核矩阵 $(K_{\phi_2 \circ \phi_0})_{ij} = K_{\phi_3}(K_{\phi_0}(x_i, x_j))$.

在这种定义的 ϕ_2 下,ODKC核对应的等价计算结构如图4所示.对于某些复杂数据集,经过一次映射之后,数据可能仍然线性不可分,有必要再进行一次映射,重构数据的几何结构,使之更好地适应分类任务.如图4所示,输入数据 \mathbf{x}_{new} 在经过两次连续的映射之后又进行了特征提取,从而形成新的特征空间(图4第四层).

对于这三种 ϕ_2 的定义,由于 k_{ϕ_0} 和 k_{ϕ_3} 都选自合理的

核,由核的闭包性质,容易证明 $k_{\phi_2 \circ \phi_0}$ 也是合理的核函数.

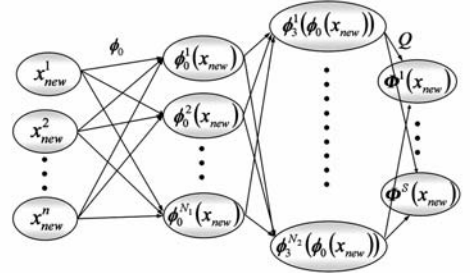


图4 在定义4的 ϕ_2 下,输入数据 \mathbf{x}_{new} 的复合特征映射 $\Phi(\cdot)$

3.4 权系数 λ_{ij} 的选择

当数据点靠近异类的类中心时,准则函数中的权矩阵 $\Lambda = (\lambda_{ij})_{m \times N}$ 起到惩罚因子的作用.其计算如下:

$$\lambda_{i,j} = \frac{H((d_{ij}^\phi)^2 - (r_i^\phi)^2) \delta_{ij}}{\sum_{j=1}^N H((d_{ij}^\phi)^2 - (r_i^\phi)^2) \delta_{ij}}$$

其中, $(d_{ij}^\phi)^2 = \|\phi(x_j) - c_i^\phi\|^2$ 是在特征映射 ϕ 对应的特征空间中,数据点 $\phi(x_j)$ 到第 i 类中心 c_i^ϕ 的距离的平方.本文采用了文献[20]中的计算公式:

$$(d_{ij}^\phi)^2 = k_\phi(x_j, x_j) - \frac{2}{N_i} \sum_{f(x_i=i)} k_\phi(x_j, x_i) + \frac{1}{N_i} \sum_{f(x_m=i)} \sum_{f(x_l=i)} k_\phi(x_l, x_m)$$

$(r_i^\phi)^2$ 是在特征映射 ϕ 对应的特征空间中,所有数据点到第 i 类中心的距离平方的平均值,计算为:

$$(r_i^\phi)^2 = -\frac{1}{N_i^2} \sum_{f(x_s=i)} \sum_{f(x_t=i)} k_\phi(x_s, x_t) + \frac{1}{N_i^2} \sum_{f(x_s=i)} k_\phi(x_s, x_s)$$

式中的 $H(x)$ 为 Heaviside 函数,即

$$H(x) = \begin{cases} x, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

4 ODKC 分类算法

归纳上述内容,本文的 ODKC 分类算法如下.

Step 1 [构造矩阵] 利用类别信息和上述各有关矩阵的定义,构造以下矩阵:

$$\delta_{m \times N}, \lambda_{m \times N}, F_{m \times N}, \bar{S}_{m \times m}, S_{m \times N}, V_{N \times m}, K_{\phi_2 \circ \phi_0}$$

Step 2 [计算广义特征值和广义特征向量] 计算式(10)的广义特征值和广义特征向量,设 $\alpha_1, \dots, \alpha_s$ 为对应于 s 个广义负特征值 $\beta_1 \leq \dots \leq \beta_s < 0$ 的广义特征向量,其中 s 为负特征值个数.

Step 3 [求最优特征映射] 根据式(11)和(12)计算 $(\Phi(x_1), \Phi(x_2), \dots, \Phi(x_N))$ 和 $\Phi(x_{test})$,ODKC核矩阵的计算由式(13)给出.

Step 4 [使用 ODKC 核矩阵进行分类] 在使用

SVM 或核 FD,核 MSE,核 KNN 等进行分类时,用 ODKC 核矩阵代替原来的基本核矩阵进行分类。

5 实验结果

为检验 ODKC 方法的有效性,本文进行了两部分实验.第一部分比较 ODKC 核与高斯核的 KTA(核与数据的匹配性)度量值.第二部分比较了使用 ODKC 核与高斯核在分类算法 SVM,FD,MSE, k -近邻(KNN)中的分类性能.在这两部分实验中,我们使用了常用的高斯核和多项式核作为 k_{ϕ_0} 和 k_{ϕ_3} 的选择.在下面各表格中,记号 P 表示多项式核,即 $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + R)^d$, R, d 为其参数; G 表示高斯核,即 $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2)$, σ 为其参数.实验中,ODKC 使用的基本核参数 R, d, σ 分别固定地选取为 4、2、33,以检验 ODKC 核无需调节核参数的自适应特性.除此之外的其它必要参数将在以下各表格的括号中给出. SVM 的参数通过交叉验证确定.此外,表格中符号 ODKC $_{p * g}$ 表示 ϕ_2 采用定义 4 的串联方式, $k_{\phi_2 * \phi_0}(\mathbf{x}, \mathbf{y}) = k_3(k_0(\mathbf{x}, \mathbf{y}))$, ϕ_0 采用高斯核, ϕ_3 采用多项式核时的最优核函数. ODKC $_{p + g}$ 表示 ϕ_2 采用定义 2 的并联方式, $k_{\phi_2 + \phi_0}(\mathbf{x}, \mathbf{y}) = k_3(\mathbf{x}, \mathbf{y}) + k_0(\mathbf{x}, \mathbf{y})$, ϕ_0 采用多项式核, ϕ_3 采用高斯核时的最优核函数. ODKC $_{p * g}$ 表示 ϕ_2 采用定义 3 的乘法方式, $k_{\phi_2 * \phi_0}(\mathbf{x}, \mathbf{y}) = k_3(\mathbf{x}, \mathbf{y}) * k_0(\mathbf{x}, \mathbf{y})$, ϕ_0 采用多项式核, ϕ_3 采用高斯核时的最优核函数. 实验选用 UCI 数据库中的五个数据集^[21]: Ionosphere, Cancer, Diabetes, Iris 和 Sonar, 其中 Iris 为三类,其余都为两类, Ionosphere 包含 351 个 34 维样本, Cancer 包含 699 个 9 维样本, Diabetes 包含 768 个 8 维样本, Iris 为包含 150 个 4 维样本, Sonar 为包含 208 个 60 维样本。

5.1 核与数据的匹配度 KTA (Kernel Target Alignment) 实验

Cristianini 等人在文献[16]中引入了核与数据的匹配度量,用来表示所采用核函数的核矩阵与目标数据的适合程度.该值在 0 和 1 之间,值越大表明所采用的核函数越匹配目标数据.匹配度定义为核矩阵 K 与标签矩阵 $\mathbf{y}\mathbf{y}^T$ 之间标准化的 Frobenius 内积,即

$$A = \frac{\langle \mathbf{K}, \mathbf{y}\mathbf{y}^T \rangle_F}{\sqrt{\langle \mathbf{K}, \mathbf{K} \rangle_F \langle \mathbf{y}\mathbf{y}^T, \mathbf{y}\mathbf{y}^T \rangle_F}}$$

其中, \mathbf{y} 为训练样本标签组成的向量. 矩阵 M 和 N 的 Frobenius 内积定义为:

$$\langle M, N \rangle_F = \sum_{i,j} M_{ij} N_{ij} = \text{tr}(M^T N)$$

为了检验 ODKC 核矩阵能否更好地适应目标数据,比较了 ODKC 核矩阵和高斯核矩阵的 KTA 值. 对数据集 Ionosphere, Cancer, Diabetes 和 Sonar, 随机选取每类数据中的 80%、50% 和 20% 作为训练样本. 每个实验重复

15 次,其 KTA 的平均值见表 1.

实验结果表明,除去 20% 训练样本下的个别情况外,ODKC 核矩阵的 KTA 值都高于高斯核矩阵的匹配值.这意味着,ODKC 核比高斯核能够更好地适应这些目标数据.

表 1 四种训练数据集的平均 KTA 值

测试数据集	训练样本比例	KTA 值			
		ODKC $_{p * g}$	ODKC $_{p + g}$	ODKC $_{p * g}$	Kg(33)
Ionosphere	20%	0.233	0.061	0.331	0.070
	50%	0.221	0.231	0.316	0.097
	80%	0.233	0.180	0.310	0.089
Sonar	20%	0.267	0.054	0.227	0.049
	50%	0.165	0.061	0.231	0.044
	80%	0.124	0.053	0.230	0.035
Cancer	20%	0.294	0.299	0.302	0.043
	50%	0.198	0.148	0.153	0.100
	80%	0.192	0.191	0.155	0.100
Diabetes	20%	0.020	0.099	0.082	0.097
	50%	0.092	0.078	0.053	0.051
	80%	0.095	0.094	0.055	0.040

5.2 分类性能实验

对于两类的问题,实验比较了采用高斯核与采用 ODKC 最优核时的 FDA、MSE 和 SVM 分类算法的性能.对于两类的数据集 Ionosphere, Cancer, Diabetes 和 Sonar, 随机选取每类数据中的 80%、50% 和 20% 样本作为训练样本,其余的用于测试.对于 Iris 数据集的多分类问题,实验比较了采用高斯核与采用 ODKC 核的 KNN 分类正确率.表中的 KNN1 表示采用高斯核、近邻数取 1 时的近邻算法. KNN3 表示采用高斯核、近邻数取 3 时的近邻算法.每个实验重复 15 次,取其平均值.限于篇幅,下面只列出了 Diabetes 和 Iris 数据集上的实验结果,见表 2 和 3.表中 SVM + ODKC $_{p * g}$ 表示采用了 ODKC $_{p * g}$ 核的 SVM 分类算法. FDA + K $_g$ 表示采用了高斯核的 FDA 分类算法,其它记号的含义类似.

表 2 Diabetes 数据集上的平均正确率 (%)

分类算法	80% 训练数据		50% 训练数据		20% 训练数据	
	训练集	测试集	训练集	测试集	训练集	测试集
SVM + ODKC $_{p + g}$	85.45	84.63	84.32	76.88	83.57	71.47
SVM + ODKC $_{p * g}$	87.56	81.75	83.64	77.87	84.27	73.54
SVM + ODKC $_{p * g}$	86.12	81.22	84.52	77.60	83.66	70.50
SVM + K $_g$ (33)	85.62	78.15	82.68	76.82	84.67	72.64
FDA + ODKC $_{p + g}$	86.46	81.65	89.74	77.34	91.03	77.76
FDA + ODKC $_{p * g}$	85.45	84.34	85.53	77.54	90.45	76.65
FDA + ODKC $_{p * g}$	88.76	82.13	89.51	77.41	90.02	75.12
FDA + K $_g$ (33)	83.16	80.04	83.85	76.56	83.65	74.95
MSE + ODKC $_{p + g}$	90.32	83.65	88.33	79.45	85.45	77.07
MSE + ODKC $_{p * g}$	88.54	84.76	87.75	82.33	86.65	76.66
MSE + ODKC $_{p * g}$	89.66	83.12	89.73	79.14	88.11	75.04
MSE + K $_g$ (33)	89.43	80.60	83.54	78.64	83.58	76.50

从实验结果中看出,对于 Iris 数据集,ODKC 分类算

法的正确率有全面改善. 对于其它数据集, ODKC 的分类算法亦好于采用高斯核的分类算法或与其相当. 需要说明的是, 这里的 ODKC 采用了固定的核参数, 并未作数据自适应的调整.

表 3 Iris 数据集上的平均正确率 (%)

分类算法	80% 训练数据	60% 训练数据
ODKC _{p+g} + KNN1	96.00	94.65
ODKC _{p*g} + KNN1	95.65	93.54
ODKC _{p^g} + KNN1	96.00	93.77
ODKC _{p+g} + KNN3	96.12	95.13
ODKC _{p*g} + KNN3	94.55	96.45
ODKC _{p^g} + KNN3	96.12	95.13
KNN1	95.12	93.69
KNN3	95.67	94.53

6 小结

本文提出了一种双核最优复合分类的构造方法 ODKC, 在统一的框架下讨论了三种不同的双核复合方式. 与传统的单核方法相比, ODKC 对复杂分类任务的适应能力得到了提高, 改善了分类效果. 另外, 由于准则函数的优化问题转化成了广义特征值问题, 所以 ODKC 有着非迭代和计算代价低的特性.

参考文献

- [1] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines [M]. Cambridge: Cambridge University Press, 2000.
- [2] Mika S, Ratsch G, et al. Fisher discriminant analysis with kernels[A]. Proceedings of the Conference on Neural Networks for Signal Processing[C]. Washington: IEEE Press, 1999. 41 – 48.
- [3] Baudat G, Anours F. Generalized discriminant analysis using a kernel approach[J]. Neural Computation, 2000, 12(10): 2385 – 2404.
- [4] Scholkopf B, Mika S, et al. Kernel PCA pattern reconstruction via approximation pre-images[A]. Proceedings of the International Conference on Artificial Neural Networks[C]. Skovde: IEEE Press, 1998. 147 – 152.
- [5] Lai P L, Fyfe C. Kernel and nonlinear canonical correlation analysis[J]. International Journal of Neural Systems, 2000, 10(5): 365 – 377.
- [6] Bach F R, Jordan M I. Kernel independent component analysis [J]. Journal of Machine Learning Research, 2002, 3(2): 1 – 48.
- [7] Ruiz A, Lopez-de P E. Nonlinear kernel-based statistical pattern analysis[J]. IEEE Trans Neural Network, 2001, 12(1): 16 – 32.
- [8] 皋军, 王士同. 基于矩阵模式的最小类内散度支持向量机 [J]. 电子学报, 2009, 37(5): 1051 – 1057.
Gao Jun, Wang Shitong. Matrix Pattern Based Minimum With-

in-Class Scatter Support Vector Machines[J]. Acta Electronica Sinica, 2009, 37(5): 1051 – 1057. (in Chinese)

- [9] 郭俊, 鲁明羽, 等. 基于混合学习框架的 SVM 反馈算法研究[J]. 电子学报, 2010, 38(9): 2101 – 2106.
Wu Jun, Lu Mingyu, et al. SVM-Feedback Scheme Within Hybrid Learning Framework for Image Retrieval[J]. Acta Electronica Sinica, 2010, 38(9): 2101 – 2106. (in Chinese)
- [10] Lanckriet G R G, Cristianini N, et al. Learning the kernel matrix with semi-definite programming [J]. Journal Machine Learning Research, 2004, 5: 27 – 72.
- [11] Bach F, Lanckriet G R G, et al. Multiple kernel learning, conic duality, and the SMO algorithm[A]. Proc 21st Int'l Conf on Machine Learning[C]. Banff: IEEE Press, 2004.
- [12] Bennett K P, Momma M, et al. Embrechts. MARK: A boosting algorithm for heterogeneous kernel models[A]. Proceedings of ACM SIGKDD[C]. Alberta: IEEE Press, 2002. 24 – 31.
- [13] Bi J, Zhang T, et al. Column-generation boosting methods for mixture of kernels[A]. Proceedings of Int'l Conf on Knowledge Discovery and Data Mining[C]. Seattle: ACM Press, 2004. 521 – 526.
- [14] Xiong H L, Swamy M N S. Optimizing the kernel in the empirical feature space [J]. IEEE transactions on neural networks, 2005, 16(2): 460 – 474.
- [15] Chen B, Liu H W, et al. Optimizing the data-dependent kernel under unified kernel optimization framework [J]. Pattern Recognition. 2007, 41(6): 2107 – 2119.
- [16] Cristianini N, Shawe-Taylor J, et al. On kernel-target alignment[A]. Advance in neural information processing systems [C]. Cambridge: MIT Press 2001.
- [17] Kingsbury N, Tay. D B H, et al. Multi-scale kernel methods for classification[A]. Proceedings of the IEEE Workshop on Machine Learning for Signal Processing[C]. Washington D C: IEEE Press, 2005. 43 – 48.
- [18] Gehler P V, Nowozin. S. Infinite Kernel Learning[R]. Germany: Max Planck Institute for Biological Cybernetics, 2008.
- [19] Bengio Y S, Yann L C, et al. Large-Scale Kernel Machines [M]. Cambridge: MIT Press, 2007.
- [20] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern analysis[M]. Cambridge: Cambridge University Press, 2004.
- [21] Blake C, Keogh E, et al. UCI Repository of Machine Learning Databases[OL]. <http://www.ics.uci.edu/mllearn>. 1998

作者简介

王 峰 男, 北京工业大学计算机学院 2009 年硕士研究生毕业. 研究方向为计算机应用.

E-mail: pr@bjut.edu.cn

张鸿宾 男, 北京工业大学计算机学院教授, 博士生导师. 主要研究方向为模式识别、图像处理与分析, 多媒体安全等.

E-mail: zhanghb@bjut.edu.cn