Kernel-kNN: 基于信息能度量的核 k-最近邻算法

刘松华 [1] 张军英 [1] 许 进 [2] 贾宏恩 [3]

摘 要 提出一种核 k 最近邻算法. 首先给出用于最近邻学习的信息能度量方法, 该方法克服了高维数据不便于用传统距离度量表示的困难, 提高了数据间类别相似性和距离的一致性. 在此基础上, 将传统的 kNN 扩展为非线性形式, 并采用半正定规划学习全局最优的度量矩阵. 算法主要特点是: 能较好地适用于高维数据, 并有效提升 kNN 的分类性能. 多个数据集的实验和分析表明, 本文的 Kernel-kNN 算法与传统的 kNN 算法比较, 在低维数据上, 分类准确率相当; 在高维数据上, 分类性能有明显提高.

关键词 距离度量, 非线性变换, k-最近邻 (k-NN), 核方法

DOI 10.3724/SP.J.1004.2010.01681

Kernel-kNN: A New kNN Algorithm Based on Informational Energy Metric

LIU Song-Hua[1] ZHANG Jun-Ying[1] XU Jin[2] JIA Hong-En[3]

Abstract This paper proposes a new algorithm named Kernel-kNN. To begin with, an approach for information energy metric is proposed, which is used to learn the nearest neighbor. This method overcomes the inconvenience for distance metric expression with high dimensional data set, and improves the consistency between the class similarity and the distance. Meanwhile, the traditional kNN is extended to an nonlinear form, and semidefinite programming is usd to learn the globally optimal metric matrix. The main characteristic of the proposed algorithm is that it is suitable for high dimensional data set, and can improve the classification performance efficiently. Experiments and analysis on many data sets have shown that Kernel-kNN can get the common performance in low dimensional data, and have a significant improvement on large scale data in high dimensions.

Key words Distance metric, nonlinear transformation, k-nearest neighbor (kNN), kernel method

kNN 根据 k 个最近邻训练样本类别对测试样本进行分类, 其分类准确率取决于样本的距离度量方法, 如欧氏距离和马氏距离. 然而距离度量容易忽略训练样本和其类别之间的统计规律, 同时当数据没有明显的距离表示的时候, 如文本、人脸识别等高维数据, kNN 则难以取得较好的分类性能[1].

近年来, 国内外学者针对提升 kNN 的分类性能提出了一些改进方法. 文献 [2−3] 通过实验验证了对输入样本进行简单的线性变换可以有效地提高 kNN 的分类性能. 文献 [4] 通过从样本中学习合适的距离度量来对人脸数据进行分类. 文献 [1, 5] 通过将相似类别的所有样本聚类来学习距离度量. 然而上述算法没有充分考虑高维数据中的高阶统计量和非线性特征. 文献 [6] 提出一种基于信息度量的方法, 并采用 Bregman 最优化方法最小化熵, 然而该方法不能保证全局最优. 文献 [7] 将传统的 kNN 扩展到核空间中, 提出核最近邻算法 (Kernel nearest neighbor, Kernel-NN), 该方法一定程度上提升了传统 kNN 的分类性能, 但由于在核空间中仍然采用欧氏距离度量, 因此不适用于提升高维数据分类效果. 文献 [8] 将 Kernel-NN 与凸包分类算法结合 (Kernel nearest neighbor convex hull, KNNCH), 增强了凸包算法的非线性分类能力. 文献 [9] 结合 kNN 与神经网络提高识别率. 文献 [10] 改进了 kNN 距离计算忽略特征权值的问题, 并用于函数回归. 文献 [11] 提出了参考样本集的最优选择方法提升 kNN 性能. 因此一个较为理想的方案是借助核方法将 kNN 推广到非线性形式, 并学习合适的度量方法, 使其不仅能充分表示数据的距离信息, 同时能有效地挖掘数据中的高阶统计量和非线性特征.

本文提出一种新的基于信息能度量的核 kNN 算法, 简称 Kernel-kNN. 算法采用数据间信息能的变化来度量其在特征空间中距离的变化; 然后给出其明确的物理意义, 来有效地解释 Kernel-kNN 算法; 最后将其规范化为半正定规划 (Semidefinite programming, SDP) 问题, 保证了全局最优.

下面首先介绍 kNN 核化的过程, 其次提出信息能度量方法, 然后给出 Kernel-kNN 的优化目标函

收稿日期 2010-03-17 录用日期 2010-08-18
Manuscript received March 17, 2010; accepted August 18, 2010
国家自然科学基金重点项目 (60933009), 国家自然科学基金 (61070137, 60702063) 资助
Supported by Key Program of National Natural Science Foundation of China (60933009), National Natural Science Foundation of China (61070137, 60702063)

1. 西安电子科技大学计算机学院 西安 710071 2. 北京大学信息科学技术学院 北京 100871 3. 西安交通大学理学院 西安 710049
1. School of Computer Science and Technology, Xidian University, Xi′an 710071 2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871 3. College of Science, Xi′an Jiaotong University, Xi′an 710049

数, 并将其与经典分类方法比较, 最后进行实验仿真, 得出在真实数据和合成数据上的实验结果.

1 Kernel-kNN

核方法能有效解决数据的线性不可分问题, 且其复杂度不依赖于数据的维数, 因此能较好地应用于高维数据. 本文将传统的 kNN 进行核化, 并学习最优的度量方法, 能在一定程度上提高其分类性能.

1.1 kNN的核化

设 $\boldsymbol{x}_{ci}$ 是输入空间 X 中第 c 类第 i 个样本, 当计算与类别无关时, 忽略类别简记为 $\boldsymbol{x}_i$. $\boldsymbol{y}_{ci}$ 为 $\boldsymbol{x}_{ci}$ 经过核映射 $\phi: X \to Y$ 后的特征空间 Y 中样本, k 为最近邻数目, N_c 表示最近邻中样本类别数目, J_c 表示最近邻中该类样本数目, 则 $c \in [1, N_c]$, $i \in [1, J_c]$.

传统的 kNN 通过输入样本的线性变换 $Y = LX$ 后能有效地提升其分类性能. 如文献 [1, 6] 中通过优化代价函数来学习距离度量矩阵, 采用的距离度量为马氏距离 M 或欧氏距离 L.

$$D_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} M (\boldsymbol{x}_i - \boldsymbol{x}_j) \tag{1}$$

其中, $M = L^{\mathrm{T}}L$.

为了挖掘样本间的非线性特征, Kernel-NN[7] 将其扩展为基于核的非线性方法, 并给出距离计算公式:

$$D(\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)) = K(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2K(\boldsymbol{x}_i, \boldsymbol{x}_j) + K(\boldsymbol{x}_j, \boldsymbol{x}_j)$$

其中, $K(\cdot, \cdot)$ 为核矩阵中元素.

然而上述核化形式不便于学习距离度量矩阵, 因此本文将其进一步改进. 由核方法知, 尽管核映射 ϕ 没有显式表示, 特征空间中样本仍然可以通过核矩阵 K 来显式表示为

$$\boldsymbol{y}_{ci} = \langle \boldsymbol{v}, \phi(\boldsymbol{x}_{ci}) \rangle = \sum_k A K(\boldsymbol{x}_k, \boldsymbol{x}_{ci}) \tag{2}$$

其中, $\boldsymbol{v}$ 为投影向量, A 为系数矩阵.

如果核映射 ϕ 取线性变换, 则由式 (1) 和式 (2) 得式 (3).

$$\begin{aligned}&(\boldsymbol{y}_i - \boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{y}_j) = \\ &(A\boldsymbol{x}_i - A\boldsymbol{x}_j)^{\mathrm{T}}(A\boldsymbol{x}_i - A\boldsymbol{x}_j) = \\ &(\boldsymbol{x}_i - \boldsymbol{x}_j)A^{\mathrm{T}}A(\boldsymbol{x}_i - \boldsymbol{x}_j) = \\ &D_M(\boldsymbol{x}_i, \boldsymbol{x}_j)\end{aligned} \tag{3}$$

如果取非线性变换, 则有式 (4) (见本页下方). 由此可知, 本文的核化便于学习最优的度量矩阵.

1.2 信息能度量方法

文献 [12] 指出高维空间中两点之间的距离可以表示为它们之间的信息量. 由于 kNN 经过核化后, 特征空间中的样本具有较高维数, 因此为了克服高维数据不便于用距离度量表示的问题, 本文用特征空间中样本间信息含量的变化来度量其距离的变化, 信息含量的变化记为信息能, 由此学习的度量称为信息能度量.

对于任一样本 $\boldsymbol{y}_{ci}$, 记为参考样本. 则参考样本邻近区域与其类别相同的, 可能成为其最近邻的样本 $\boldsymbol{y}_{cj}$, 记为目标近邻; 与其类别不同的近邻 $\boldsymbol{y}_{pl}$, 记为入侵近邻, 且 $c \neq p$. 对于 $\boldsymbol{y}_{ci}$, 其信息能计算包括两项: 第一项为 $\boldsymbol{y}_{cj}$ 的同类信息能, 由于属于同一类, 因此其值越大, 对分类性能提升影响越大, 记为 E_c; 第二项为 $\boldsymbol{y}_{pl}$ 的异类信息能, 且属于不同类别, 因此其值越小, 对分类影响越小, 记为 $E_{p\neq c}$. 分别计算如下

$$E_c(\boldsymbol{y}_{ci}) = \sum_{j=1}^{J_c} G(\boldsymbol{y}_{cj} - \boldsymbol{y}_{ci}, 2\sigma^2 I) \tag{5}$$

$$E_{p\neq c}(\boldsymbol{y}_{ci}) = \sum_{p=1}^{N_c} \sum_{l=1}^{J_p} G(\boldsymbol{y}_{pl} - \boldsymbol{y}_{ci}, 2\sigma^2 I) \tag{6}$$

其中, G 为核密度估计函数[13−14], 本文采用高斯核函数, σ 为待定参数, I 为单位矩阵.

则任一参考样本 $\boldsymbol{y}_{ci}$ 的信息能为

$$E = aE_c - (1-a)E_{p\neq c} \tag{7}$$

其中, $a = \left[\left(1 - \frac{J_c}{k+1}\right)^2 + \sum_{\substack{p=1 \\ p\neq c}}^{N_c} \left(\frac{J_p}{k+1}\right)^2\right]$, k 为参考样本近邻数, a 为目标近邻和入侵近邻对信息能计算的影响参数, 其中第一项表示对于第 c 类参考样本, 其他所有样本先验概率的影响; 第二项表示其他各类样本各自先验概率的影响.

$$\begin{aligned}(\boldsymbol{y}_i - \boldsymbol{y}_j)^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{y}_j) = &\left(\sum_k K(\boldsymbol{x}_k, \boldsymbol{x}_i) - \sum_l K(\boldsymbol{x}_l, \boldsymbol{x}_j)\right)^{\mathrm{T}} A^{\mathrm{T}}A \left(\sum_k K(\boldsymbol{x}_k, \boldsymbol{x}_i) - \sum_l K(\boldsymbol{x}_l, \boldsymbol{x}_j)\right) = \\ &\left(\sum_k K(\boldsymbol{x}_k, \boldsymbol{x}_i) - \sum_l K(\boldsymbol{x}_l, \boldsymbol{x}_j)\right)^{\mathrm{T}} M \left(\sum_k K(\boldsymbol{x}_k, \boldsymbol{x}_i) - \sum_l K(\boldsymbol{x}_l, \boldsymbol{x}_j)\right) = \\ &D_M(K(\boldsymbol{x}_k, \boldsymbol{x}_i), K(\boldsymbol{x}_l, \boldsymbol{x}_j))\end{aligned} \tag{4}$$

由于高维数据通常不满足距离度量公理中的三角不等式, 因此直接用距离度量会导致数据类别相似和距离相近的不一致性, 属于非距离度量问题[15]. 而本文提出的信息能度量通过将样本映射到特征空间, 对特征空间中样本采用信息能度量, 满足距离度量公理, 且信息能等价于互信息的计算[13], 下面给出证明.

定理 1. 在特征空间 Y 上的信息能度量 E, 对于所有 Y 中的 $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$, E 满足距离度量公理.

证明. 1) 不失一般性, 设 $\boldsymbol{x}$ 属于第 c 类, $\boldsymbol{y}$ 属于第 p 类, 样本 $\boldsymbol{x}$, $\boldsymbol{y}$ 的信息能为

$$E(\boldsymbol{x}, \boldsymbol{y}) = E(\boldsymbol{x}) + E(\boldsymbol{y})$$

由式 (7) 知

$$\begin{cases} E(\boldsymbol{x}) = aE_c(\boldsymbol{x}) - (1-a)E_{p\neq c}(\boldsymbol{x}) \\ E(\boldsymbol{y}) = a'E_p(\boldsymbol{y}) - (1-a')E_{c\neq p}(\boldsymbol{y}) \end{cases}$$

其中, 参数 a, a' 根据式 (7) 中的参数计算方法获取, 由于只考虑样本 $\boldsymbol{x}$, $\boldsymbol{y}$, 因此 $J_c = J_p = 1$, $k = 1$, 得 $a = a' = 1/2$. 同时根据式 (5) 和式 (6) 得 $E_c = E_p = 1$, $E_{p\neq c} = G(\boldsymbol{y} - \boldsymbol{x}, 2\sigma^2 I)$, $E_{c\neq p} = G(\boldsymbol{x} - \boldsymbol{y}, 2\sigma^2 I)$. 由文献 [13−14] 中高斯核函数性质知 $E_{p\neq c} = E_{c\neq p} = G(\boldsymbol{x} - \boldsymbol{y}, 2\sigma^2 I)$ 且 $0 \leq G(\boldsymbol{x} - \boldsymbol{y}, 2\sigma^2 I) \leq 1$, 因此信息能

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{E_{p\neq c}(\boldsymbol{x}) + E_{c\neq p}(\boldsymbol{y})}{2} = \\ 1 - G(\boldsymbol{x} - \boldsymbol{y}, 2\sigma^2 I) = \\ E(\boldsymbol{y}, \boldsymbol{x}) \geq 0 \end{aligned}$$

即信息能度量 E 满足非负性和对称性.

2) 设 $\boldsymbol{x}$ 属于第 c 类, 则 $E(\boldsymbol{x}, \boldsymbol{x}) = 2E(\boldsymbol{x})$. 根据式 (5) 和式 (6) 知 $E(\boldsymbol{x}) = aE_c(\boldsymbol{x}) - (1-a)E_p(\boldsymbol{x})$, 与 1) 中分析相同, 得 $a = 0, E_c = 1, E_p = 0$. 得 $E(\boldsymbol{x}, \boldsymbol{x}) = 0$, 满足同一性.

3) 不失一般性, 设 $\boldsymbol{x}, \boldsymbol{y}$ 均属于第 c 类, z 属于第 p 类, 由于考虑 3 个样本, 因此 $J_c = 2, J_p = 1$,k= 2. 将 $G(* - *, 2\sigma^2 I)$ 简记为 $G(* - *)$, 得式 (8).

$$\begin{cases} E(\boldsymbol{x}, \boldsymbol{z}) = \dfrac{2G(\boldsymbol{x}-\boldsymbol{y})}{9} - \dfrac{8G(\boldsymbol{x}-\boldsymbol{z})}{9} - \dfrac{G(\boldsymbol{y}-\boldsymbol{z})}{9} + \dfrac{11}{9} \\ E(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{4G(\boldsymbol{x}-\boldsymbol{y})}{9} - \dfrac{7G(\boldsymbol{x}-\boldsymbol{z})}{9} - \dfrac{7G(\boldsymbol{y}-\boldsymbol{z})}{9} + \dfrac{4}{9} \\ E(\boldsymbol{y}, \boldsymbol{z}) = \dfrac{7G(\boldsymbol{x}-\boldsymbol{y})}{9} - \dfrac{G(\boldsymbol{x}-\boldsymbol{z})}{9} - \dfrac{3G(\boldsymbol{y}-\boldsymbol{z})}{9} + \dfrac{15}{9} \end{cases} \tag{8}$$

由式 (8) 知

$$E(\boldsymbol{x}, \boldsymbol{z}) \leq E(\boldsymbol{x}, \boldsymbol{y}) + E(\boldsymbol{y}, \boldsymbol{z})$$

即 E 满足三角不等式.

综上可知度量 E 满足距离度量公理. □

1.3 Kernel-kNN 算法实现及其物理意义

为有效解释本文算法的目的, 采用文献 [13] 的记法. 将样本和其近邻看作一个信息势能场, 样本看作信息微粒. 在样本的输入空间和特征空间, 由于核映射对样本进行的变换, 会导致特征空间中样本相对输入空间 “移动”, 因此, 相对输入空间, 在特征空间内, 对于参考样本微粒 $\boldsymbol{y}_{ci}$, 同类的微粒 $\boldsymbol{y}_{cj}$ 如果向其靠近, 则会增加 $\boldsymbol{y}_{cj}$ 的信息能. 不同类的微粒 $\boldsymbol{y}_{pj}$ 如果远离 $\boldsymbol{y}_{ci}$, 则会减少势能场中 $\boldsymbol{y}_{pj}$ 的信息能. 根据 kNN 分类规则: 样本的邻近区域应该包含尽可能多的目标近邻, 同时入侵近邻应该尽可能少. 因此本文信息能代价函数设置两项: 第一项增加样本与其目标近邻的信息能 $\xi_{\text{pull}}(A)$, 使其靠近该样本, 其中, A 为式 (2) 中的系数矩阵; 第二项减少与其入侵近邻的信息能 $\xi_{\text{push}}(A)$, 使其远离该样本. 为保证其满足 SDP 条件, 设 $M = A^{\mathrm{T}}A$, 则信息量代价函数为

$$\xi(M) = \xi_{\text{pull}}(M) + \xi_{\text{push}}(M) + E_0 \tag{9}$$

其中, $\xi_{\text{pull}}(M) = \sum_{c=1}^{N_c} aE_c$, 可由式 (5) 计算. 由式 (6) 得 $\xi_{\text{push}}(M) = \sum_{p=1}^{N_c}(a-1)E_{p\neq c}$, E_0 可以解释为间隔量 (Margin), 表示入侵近邻微粒从最远目标近邻的势能位置移动到安全距离外信息能的变化量. 因此, 将核方法引入传统 kNN, 并采用信息能度量具有较好的物理意义.

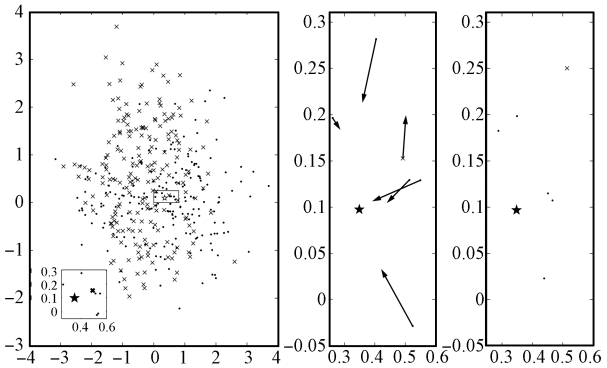
由上可知, Kernel-kNN 解决如下最优化问题

$$\begin{aligned} &\max_{M}(1-\mu)\xi_{\text{pull}}(M) + \mu\xi_{\text{push}}(M) + E_0 \\ &\text{s.t.} \begin{cases} 1)\ \xi_{\text{pull}}(M) + E_0 \leq \xi_{\text{push}}(M) \\ 2)\ M \geq 0 \end{cases} \end{aligned} \tag{10}$$

其中, μ 调节同类样本与不同类样本信息势能的变化, 可以通过交叉验证获取. 条件 1) 保证入侵近邻的信息能增加量大于目标近邻信息能增加量与间隔量之和; 条件 2) 保证获取的系数矩阵 M 满足 SDP 条件.

通过求解式 (10) 的 SDP 问题, 可以得到式 (2) 中的系数矩阵 A 或 M, 其中, $M = A^{\mathrm{T}}A$. 图 1 给出了 Kernel-kNN 在合成数据中的实验效果.

图 1 中实验采用合成的 2 类 3 维数据, 核映射 ϕ 取线性变换. 第一类从中心在 $(1, 0, 0)$ 和 $(-1, 0, 0)$ 的双峰高斯分布中随机抽取 200 个样本, 方差为 0.9. 第二类从中心在 (0, 1, 0) 和 (0, −1, 0) 的双峰高斯分布中随机抽取 200 个样本, 方差为 0.3. 图 1 (a) 为原始样本在二维空间中的分布, 图中左下角矩形图包括样本的局部放大显示, 其中, 参考样本用 “⋆” 表示, 目标近邻用 “·” 表示, 入侵近邻用 “×” 表示.



(a) 原始合成数据 (b) 学习前输入空间数据 (c) 学习后特征空间数据

(a) The original synthetic data (b) The data of input space before learning (c) The data of feature space after learning

图 1 Kernel-kNN 在合成数据上的实验效果

Fig. 1 Experimental results of Kernel-kNN on synthetic data

对于参考样本, 其目标近邻将移向该微粒, 因此作用方向指向参考样本, 而入侵近邻将远离参考样本, 因此作用方向偏离参考样本, 如图 1 (b) 所示. 对参考样本信息能求偏导可以获取对该微粒的信息作用力方向.

图 1 (c) 显示样本经过 Kernel-kNN 系数矩阵变换后获取的相对位置, 由图可知入侵近邻微粒从最远目标近邻的势能位置移动到了安全距离外, 有效地保证了样本的可分性.

1.4 Kernel-kNN 与 LMNN[3] 及 Kernel-NN[14] 的比较

LMNN 为线性方法, 通过学习最优的马氏度量矩阵来提升 kNN 的分类性能, 分析 Kernel-kNN 的代价函数可知文献 [1] 中的优化问题是本文算法的特例, 证明如下.

定理 2. 将本文的核映射取为线性变换, 则文献 [1] 中的优化问题可由 Kernel-kNN 的优化问题式 (10) 得出.

证明. 文献 [1] 中的代价函数如式 (11) 所示. 其中, $j \to i$ 表示样本 $\boldsymbol{x}_j$ 为样本 $\boldsymbol{x}_i$ 的目标近邻, 且属于同一类别. c_{il} 表示当样本 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_l$ 的类别相同时值为 0, 类别不同时值为 1. 式 (11) 中参数 μ 的变化对结果不敏感, 因此忽略. 且第二项处理类别不同时的样本, 因此对式 (11) 采用本文的记法, 其中 $p, q \in [1, N_c]$, 且 $p \neq q$.

本文的代价函数做如下假设可以近似得到式 (12) 所示的代价函数.

1) 间隔量 $E_0 = \text{num}(l)$, 其中, l 表示入侵近邻下标, 其数目为 $\text{num}(l)$. 忽略 μ (对实验结果不敏感);

2) 式 (10) 的代价函数中, 设核映射 ϕ 取线性变换, 即 $Y = AX$, 设 $M = A^{\mathrm{T}}A$. 式 (2) 和式 (3) 中高斯函数 G 取线性变换 $G(Y) = Y^{\mathrm{T}}Y$.

根据上述假设, 本文代价函数可以计算如式 (12) 所示. 式中高斯核函数计算转化为

$$\begin{aligned} G(\boldsymbol{y}_{pi} - \boldsymbol{y}_{pj}) = &(\boldsymbol{y}_{pi} - \boldsymbol{y}_{pj})^{\mathrm{T}}(\boldsymbol{y}_{pi} - \boldsymbol{y}_{pj}) = \\ &(\boldsymbol{x}_{pi} - \boldsymbol{x}_{pj})^{\mathrm{T}} M (\boldsymbol{x}_{pi} - \boldsymbol{x}_{pj}) = \\ &D_M(\boldsymbol{x}_{pi} - \boldsymbol{x}_{pj}) \end{aligned}$$

由此可知, 式 (12) 计算可以转化为式 (13), 其中, $b = (a-1)/\text{num}(i \times l)$, $c = (1-a)/\text{num}(i)$, $a = [(1 - J_c/(k+1))^2 + \sum_{\substack{p=1 \\ p \neq c}}^{N_c} (J_p/(k+1))^2]$.

观察 $\xi'(M)$ 与 $\xi(M)$ 知 b, c 都取 1 时两式相同. □

$$\begin{aligned} \xi'(M) = &(1-\mu)\xi'_{\text{pull}}(M) + \mu\xi'_{\text{push}}(M) = \\ &(1-\mu)\sum_{j \to i} D_M(\boldsymbol{x}_i, \boldsymbol{x}_j) + \mu \sum_{i,j \to i}\sum_{l}(1 - c_{il})\left[1 + D_M(\boldsymbol{x}_i, \boldsymbol{x}_j) - D_M(\boldsymbol{x}_i - \boldsymbol{x}_l)\right] = \\ &\sum_{j \to i} D_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{pj}) + \sum_{i,j \to i}\sum_{l}\left[1 + D_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{pj}) - D_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{ql})\right] \end{aligned} \tag{11}$$

$$\begin{aligned} \xi(M) = &\xi_{\text{pull}}(M) + \xi_{\text{push}}(M) + E_0 = aE_p + (a-1)E_{c \neq p} + E_0 = \\ &a\sum_{j \to i} G(\boldsymbol{y}_{pi} - \boldsymbol{y}_{pj}) + (a-1)\sum_{l} G(\boldsymbol{y}_{ql} - \boldsymbol{y}_{pi}) + E_0 \end{aligned} \tag{12}$$

$$\begin{aligned} \xi(M) = &a\sum_{j \to i} D_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{pj}) + (a-1)\sum_{l} D_M(\boldsymbol{x}_{ql}, \boldsymbol{x}_{pi}) + E_0 = \\ &\sum_{j \to i} D_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{pj}) + \sum_{i,j \to i}\sum_{l}\left[1 + bD_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{pj}) - cD_M(\boldsymbol{x}_{pi}, \boldsymbol{x}_{ql})\right] \end{aligned} \tag{13}$$

不同于 LMNN, 核函数的引入使 Kernel-kNN 能处理非线性变换问题, 并能有效地提取高维空间中样本的非线性结构, 增强其类别可分性.

与 Kernel-NN 比较, Kernel-kNN 有如下特点:

1) 用信息能度量学习合适的高维数据度量矩阵, 能提取高阶统计量;

2) 克服了高维数据不便于用距离度量的困难, 用信息能度量代替 Kernel-NN 中使用的欧氏距离度量;

3) 通过度量矩阵学习, 使得同类样本靠近, 异类样本远离, 更符合 kNN 的分类规则, 能有效提升分类性能.

2　实验与分析

本文在维数不同、样本数目差别较大的 9 个数据集上进行分类性能测试. 实验包括两部分: 1) 小规模数据集测试; 2) 大规模数据集测试, 包括了图像、语音、文本等高维数据. 表 1 给出了数据集的相关属性. 分类性能测试算法包括线性的 kNN、LMNN、ITML[6], 以及基于核的非线性方法 Kernel-NN、Multiclass SVM[16]、KNNCH、Kernel-kNN. 所有的 kNN 算法近邻数目均取 3, 核参数参照 Multiclass SVM 中的方法选取.

2.1　小规模数据集测试结果分析

数据集 Bal、Wine、Iris 均为 UCI 标准数据集[17], 其训练样本数目均小于 500, 且包含 3 类. 对于没有测试样本的数据, 从输入样本中抽取 70 % 作为训练样本, 30 % 作为测试样本. 表 2 中小规模数据集上的分类错误率均为 100 次实验的平均值. 人脸数据 oFaces 从人脸数据库获取[18], 为了与文献 [1] 比较, 采用相同的处理方法, 用主成分分量分析 (Principle component analysis, PCA) 进行降维. 在这些数据库上, Kernel-kNN 有效地提升了传统 kNN 的分类性能. 由表 2 知 Kernel-kNN 接近 Multiclass SVM 的分类性能, 在 Wine 上优于 SVM, 在大部分数据上优于 Kernel-NN. KNNCH[8] 由于只在 oFaces 数据上进行测试, 因此表中仅列出该数据上的测试错误率, 并选 KNNCH 的最小错误率. Multiclass SVM 的结果均为文献 [1] 中的测试误差.

以人脸识别数据 oFaces 为例, 图 2 所示为原始人脸数据在二维空间内的显示. oFaces 数据库包括 40 个类, 每个人有 10 个不同的表情, 从每个人的 10 个样本中随机抽取 7 个作为训练样本, 3 个作为测试样本. 图 3 所示为 oFaces 中抽取的测试样本在输入空间内 30 像素 × 30 像素方格中的二维显示. 由图 3 可知, 第一类的 3 个测试样本分别位于图中坐标 (12, 19)、(13, 20)、(15, 14) 处, 根据欧氏距离, 前两个样本距离较近, 而第三个样本则远远偏离了同类样本, 因此不适合采用欧氏距离.

由图 4 可知, 经过 Kernel-kNN 学习后, 图 3 中

表 1　标准数据集

Table 1　Standard datasets

	大规模数据集					小规模数据集			
	Mnist	Letters	20news	Isolet	yFaces	Bal	oFaces	Wine	Iris
样本数目	70 000	20 000	18 827	7 797	2 414	625	400	178	150
特征维数	784	16	20 000	617	8 064	4	4 096	13	4
降维数目	164	16	200	172	300	4	200	13	4
训练样本	2 000	2 000	2 000	1 976	1 690	465	280	124	105
测试样本	2 000	1 600	1 400	1 559	724	160	120	54	45
类别数目	10	26	20	26	38	3	40	3	3

表 2　标准数据集上训练/测试错误率

Table 2　Train/test error rates on standard datasets

	大规模数据集					小规模数据集			
	Mnist	Letters	20news	Isolet	yFaces	Bal	oFaces	Wine	Iris
kNN	10.65/8.45	9.90/11.81	49.40/54.64	18.78/10.26	37.04/29.19	20.00/15.00	5.80/6.03	24.19/29.63	0.95/6.67
Kernel-NN	7.60/6.25	6.45/6.25	37.40/25.72	0.87/5.24	15.21/12.13	13.25/6.45	3.84/3.11	3.10/2.67	0.85/4.85
LMNN	7.25/6.15	6.15/5.56	33.21/23.25	0.05/5.13	15.50/10.11	13.12/6.25	4.72/3.16	4.03/3.70	0.95/5.79
ITML	6.75/5.60	6.05/4.88	28.25/20.64	0.61/5.84	27.34/16.49	14.02/6.67	4.26/3.20	3.09/2.57	0.65/4.67
Multiclass SVM	−/1.20	−/3.21	−/8.04	−/3.40	−/15.22	−/1.92	−/1.90	−/22.24	−/3.45
KNNCH	−/−	−/−	−/−	−/−	−/−	−/−	−/2.50	−/−	−/−
Kernel-kNN	6.60/5.60	4.45/4.25	16.50/18.93	0.96/4.94	15.56/10.47	13.12/6.25	3.11/2.65	2.10/1.85	0.95/5.68

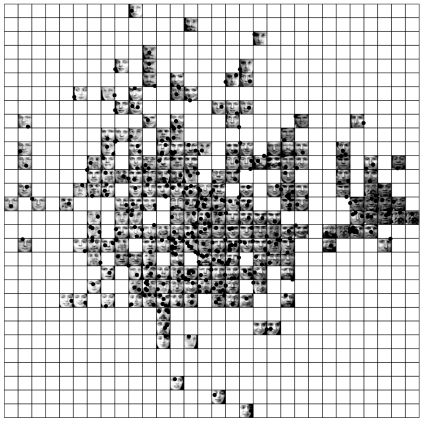


图 2 输入空间人脸数据: oFaces

Fig. 2 The human face data of input space: oFaces

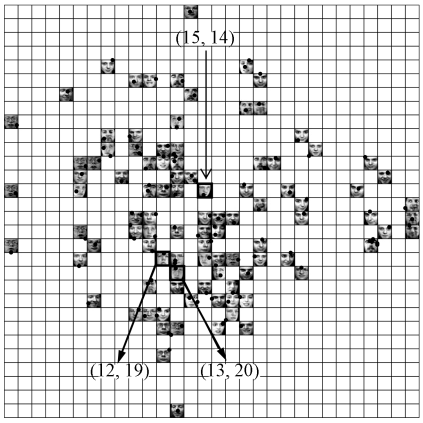


图 3 Kernel-kNN 学习前输入空间的人脸测试数据

Fig. 3 Face test data before Kernel-kNN learning of input space

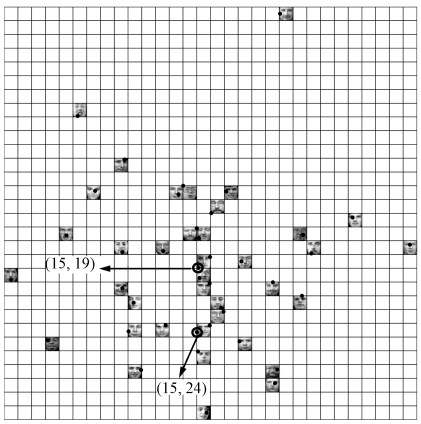


图 4 Kernel-kNN 学习后特征空间的人脸测试数据

Fig. 4 Face test data after Kernel-kNN learning of feature space

给出坐标的第一类 3 个测试数据聚集度 (与第 5 类聚于方格图 4 (15, 24) 中) 较高, 体现了同类样本相似性和距离的一致性. 同时测试数据中共有 36 类数据均显示了较好的分类效果, 除上述两类, 另外两类 (第 6 类与第 18 类) 聚集位置见图 4 方格 (15, 19).

同时与传统 kNN 比较, 本文算法引入核方法之后能有效提取高维数据中的非线性特征, 有益于提升分类性能, 图 5 所示为传统 kNN 学习后的测试数据显示. 从图中可知, 第一类人脸测试数据最后显示坐标分别为 (10, 22), (11, 22), (12, 21). 然而其他不同类别之间的人脸数据较为分散, 图 4 与图 5 的比较体现了 Kernel-kNN 在高维数据中进行分类的有效性.

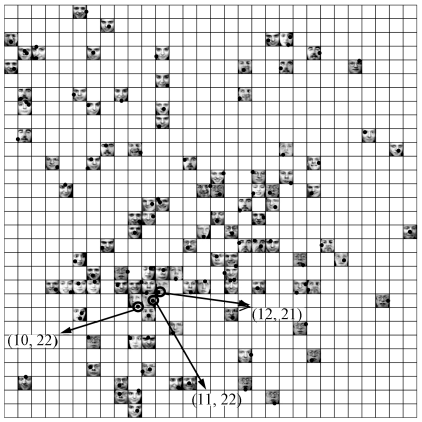


图 5 kNN 学习后特征空间的人脸测试数据

Fig. 5 Face test data after kNN learning of feature space

2.2 大规模数据集测试结果分析

为了进一步验证本文算法在高维数据上的分类性能, 采用高维的手写字体 Mnist[19]、字符识别 Letter[17]、文本分类 20news[20]、语音字符识别 Isolet[17]、人脸数据 yFaces[18]. 对于数据集中有训练数据和测试数据的样本, 随机从训练样本中抽取小于 2 000 个作为训练数据, 从测试样本中抽取小于 2 000 个作为测试数据, 除 Letter 外, 其余均采用 PCA 进行降维, 降维维数和抽取样本数据见表 1, 分类训练/测试性能见表 2, 表中训练/测试错误率均为 10 次实验的平均值.

Multiclass SVM 的实验结果由于采用文献 [1] 的样本数目进行训练, 在各数据集中均获得了较好的分类性能, 而 Kernel-kNN 在核矩阵的运算过程中, 受内存限制, 所有样本数据集均抽取小于 2 000 个, 因此分类性能相对较差.

在 Mnist 手写体识别中, 核函数采用与 Multiclass SVM 相同的线性变换, 实验结果表明 Kernel-kNN 尽管采用了部分随机数据测试, 其错误率仅为 5.6 %. 在其他高维数据中采用与 Multiclass SVM 相同的高斯核函数, 也取得了较好的分类性能. 其中, 在 Letter、Isolet 数据集上的性能与 Multiclass SVM 接近; 在 20news 数据集上, 尽管无法与 Mul-

ticlass SVM 比较, 但相对传统 kNN, Kernel-kNN 在很大程度上提升了 kNN 的分类性能; 在 yFaces 人脸识别数据集上, Kernel-kNN 的分类性能超过了 Multiclass SVM; 在大部分数据上, Kernel-kNN 的分类性能均优于 Kernel-NN.

因此 Kernel-kNN 能较好地适用于没有明显距离表示的高维数据.

在上述实验中, 本文考虑了降维维数对分类性能的影响, 以 Isolet 数据为例, 用 PCA 对数据进行降维, 从 17 维到 172 维, 分别考虑维数对分类性能的影响, 如图 6 所示. 图中分别对 kNN、Kernel-NN、LMNN、ITML、Kernel-kNN 这 5 种分类方法进行测试, 横坐标为降维维数, 纵坐标为训练错误率. 由图可知 Kernel-kNN 的训练准确率优于传统 kNN, 并与其他 4 种分类方法相近.

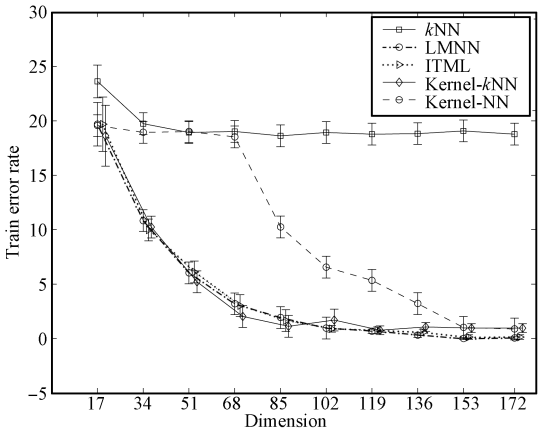


图 6　维数对训练错误率的影响

Fig. 6　Impact of dimension over train error rate

图 7 为降维维数对测试错误率的影响. 由此可知, Kernel-kNN 最大的特点是在较低维的时候能获取较高的分类准确率, 在维数为 68 的时候, 测试错误率最先到达最低点, 随着维数增加, 分类方法的测试错误率均有下降. 但超过一定维数的时候, 对分类性能提升影响较小.

本文还考虑了其他影响因素, 如在核函数的选取上, 除了 Mnist 采用线性核函数能获取较好的结果, 在其他数据集上高斯核函数均能取得较优的分类性能. 此外目标优化函数式 (10) 中的参数 μ 采用交叉验证, 其取值对分类性能影响较小, 因此设为 0.5.

最后, 分析传统 kNN 知, 其时间复杂度为 $\mathrm{O}(Nd)$, N 为样本数目, d 为样本维数. Kernel-NN 只改变了其距离计算方法, 因此时间复杂度与传统 kNN 相同. 而本文算法引入半正定规划对近邻进行优化, 且仅对核矩阵进行计算, 时间复杂度为 $\mathrm{O}(N^2)$. 因此当样本数目增加时, 将成为算法性能[21] 提升的瓶颈.

综上所述, Kernel-kNN 能在一定程度上提升传统 kNN 的分类性能.

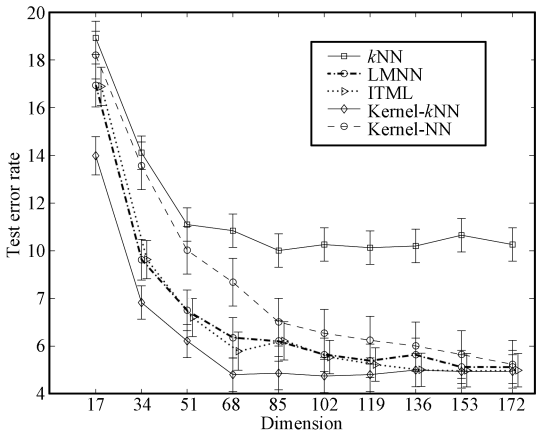


图 7　维数对测试错误率的影响

Fig. 7　Impact of dimension over test error rate

3　结束语

本文以信息能最大化为准则, 引入核方法, 利用信息能度量方法学习最优的度量矩阵, 使得 Kernel-kNN 具有较好的分类性能. 主要工作如下: 1) 信息能度量的提出, 实现了算法核化的关键步骤, 并能有效地提取数据中的高阶统计量; 2) 与 Kernel-NN 相比, 本文的信息能度量有效解决了高维数据不便于用欧氏距离表示的困难, 并在学习过程中保证了同类样本的聚集, 有利于分类性能的提升; 3) SDP 规划问题能保证全局最优, 对初始系数矩阵不敏感.

本文最后使用合成数据和真实数据的实验证明了提出算法的合理性. 尽管取得了一些有益的结果, 但核矩阵运算量问题、高维数据的降维方法等仍然需要进一步深入研究和探讨.

致谢

感谢 Weinberger K. Q. 对本文仿真实验程序和数据提供帮助.

References

1 Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 2009, **10**: 207−244

2 Goldberger J, Roweis S, Hinton G, Salakhutdinov R. Neighbourhood components analysis. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2004. 513−520

3 Torresani L, Lee K C. Large margin component analysis. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2007. 1385−1392

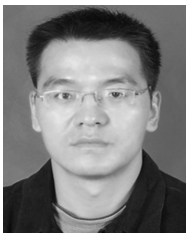
4 Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In:

Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 539−546
5 Globerson A, Roweis S T. Metric learning by collapsing classes. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2006. 451−458
6 Davis J V, Kulis B, Jain P, Sra S, Dhillon I S. Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning. Corvallis, USA: ACM, 2007. 209−216
7 Yu K, Ji L, Zhang X G. Kernel nearest-neighbor algorithm. *Neural Processing Letters*, 2002, **15**(2): 147−156
8 Zhou Xiao-Fei, Yang Jing-Yu, Jiang Wen-Han. Kernel nearest neighbor convex hull classification algorithm. *Journal of Image and Graphics*, 2007, **12**(7): 1209−1213
(周晓飞, 杨静宇, 姜文瀚. 核最近邻凸包分类算法. 中国图象图形学报, 2007, **12**(7): 1209−1213)
9 Hao Hong-Wei, Jiang Rong-Rong. Training sample selection method for neural networks based on nearest neighbor rule. *Acta Automatica Sinica*, 2007, **33**(12): 1247−1251
(郝红卫, 蒋蓉蓉. 基于最近邻规则的神经网络训练样本选择方法. 自动化学报, 2007, **33**(12): 1247−1251)
10 Ye Tao, Zhu Xue-Feng, Li Xiang-Yang, Shi Bu-Hai. Soft sensor modeling based on a modified k-nearest neighbor regression algorithm. *Acta Automatica Sinica*, 2007, **33**(9): 996−999
(叶涛, 朱学峰, 李向阳, 史步海. 基于改进 k-最近邻回归算法的软测量建模. 自动化学报, 2007, **33**(9): 996−999)
11 Zhang Hong-Bin, Sun Guang-Yu. Optimal selection of reference subset for nearest neighbor classification. *Acta Electronica Sinica*, 2000, **28**(11): 16−21
(张鸿宾, 孙广煜. 近邻法参考样本集的最优选择. 电子学报, 2000, **28**(11): 16−21)
12 Amari Shun-ichi, Nagaoka H. *Methods of Information Geometry (Translations of Mathematical Monographs)*. New Orleans: American Mathematical Society, 2000
13 Torkkola K. Feature extraction by non-parametric mutual information maximization. *The Journal of Machine Learning Research,* 2003, **3**: 1415−1438
14 Xu Dong-Bin, Huang Lei, Liu Chang-Ping. Adaptive kernel density estimation for motion detection. *Acta Automatica Sinica*, 2009, **35**(4): 379−385
(徐东彬, 黄磊, 刘昌平. 自适应核密度估计运动检测方法. 自动化学报, 2009, **35**(4): 379−385)
15 Zhang Y, Zhou Z H. Non-metric label propagation. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, USA: Morgan Kaufmann Publishers, 2009. 1357−1362
16 Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. *The Journal of Machine Learning Research*, 2002, **2**: 265−292
17 UCI machine learning repository [Online], available: http://archive.ics.uci.edu/ml/, March 10, 2009
18 Face data [Online], available: http://www.uk.research.att.com/facedatabase.html, March 10, 2009
19 LeCun Y, Cortes C. Mnist data [Online], available: http://yann.lecun.com/exdb/mnist/, March 10, 2009
20 Newsgroups data [Online], available: http://people.csail.mit.edu/jrennie/20Newsgroups, March 10, 2009
21 Arya S, Malamatos T, Mount D M. Space-time tradeoffs for approximate nearest neighbor searching. *Journal of the Association for Computing Machinery*, 2009, **57**(1): 1−54

刘松华 西安电子科技大学计算机学院博士研究生. 主要研究方向为机器学习和智能信息处理. 本文通信作者.
E-mail: sooh.liu@gmail.com
(**LIU Song-Hua** Ph. D. candidate at the School of Computer Science and Technology, Xidian University. His research interest covers machine learning and intelligent information processing. Corresponding author of this paper.)

张军英 西安电子科技大学计算机学院教授. 主要研究方向为人工神经网络、图像处理、模式识别、优化、智能信息处理与生物信息学.
E-mail: jyzhang@mail.xidian.edu.cn
(**ZHANG Jun-Ying** Professor at the School of Computer Science and Technology, Xidian University. Her research interest covers neural networks, image processing, pattern recognition, optimization, intelligent information processing, and bioinformatics.)

许 进 北京大学信息科学技术学院教授. 主要研究方向为生物计算机、生物信息处理、图论与优化计算.
E-mail: jxu@pku.edu.cn
(**XU Jin** Professor at the School of Electronics Engineering and Computer Science, Peking University. His research interest covers biological computer, bioinformatics, and graph theory and optimization.)



贾宏恩 西安交通大学理学院博士研究生. 主要研究方向为微分方程数值解.
E-mail: jiahongen@yahoo.com.cn
(**JIA Hong-En** Ph. D. candidate at the College of Science, Xi′an Jiaotong University. His research interest covers numerical solution of differential equations.)