

基于概率测度支持向量机的静态手写数字识别方法

方 向, 陈思佳, 贾 颖

(山东工商学院 计算机部, 烟台 264005)

摘 要: 提出了基于概率测度的支持向量机算法, 它采用概率分布作为均值嵌入构造再生希尔伯特空间, 为了能够直接采用任何标准的基于核的学习技术, 又构造了支持向量机的一般形式, 称为基于概率测度的支持向量机 (PM-SVM). 通过在 MNIST 数据库构建的虚拟样本进行实验, 证明了该算法在识别率和时间消耗上更为有效.

关键词: 静态手写数字识别; 支持向量机; 统计学习理论

中图分类号: TP319

文献标识码: A

文章编号: 1000-7180(2015)04-0107-04

Off-line Handwritten Digit Recognition by SVM Based on Probability Measure

FANG Xiang, CHEN Si-Jia, JIA Ying

(Department of Computer Science, Shandong Institute of Business and Technology, Yantai 264005, China)

Abstract: This paper proposes a support vector machine (SVM) algorithm based on probability measure. It employs these probability distributions as embeddings to reproduce kernel Hilbert space (RKHS). In order to reuse many standard kernel-based learning techniques in straightforward fashion, we construct the general form of support vector machine (SVM) called support vector machine based on probability measure (PM-SVM). We set up a virtual database by invariant transformation of the image through the MNIST database, and apply our algorithm to this database. The experimental results demonstrate the effectiveness of this algorithm in rate and efficiency to handwritten digit recognition.

Key words: handwritten digit recognition; SVM; statistical learning theory

1 引言

近年来, 人们对手写数字识别研究表现出极大的兴趣^[1]. 作为图像处理与模式识别领域的一个具有较高实用价值的重要问题, 手写数字识别可以分为动态(联机)手写数字识别和静态(脱机)手写数字识别两种. 由于动态识别可以考虑数字输入过程中的笔顺和笔画数, 因此相比静态识别更容易, 研究成果也更成熟. 静态手写数字识别只能提供二维图像信息不能提供时序信息, 因此, 到目前为止, 静态手写数字的识别还是特别有难度的^[2].

目前, 在静态手写数字识别中常用的分类算法有: K-最近邻算法、BP 神经网络、支持向量机等^[3].

K-最近邻分类器的识别率相对较低^[4]; BP 神经网络的前馈性使得它的识别率相对较高, 但需要大量的训练数据, 速度较慢^[5]; 支持向量机是利用凸优化理论来训练分类器, 它的训练过程比神经网络更简单, 在测试阶段, 它的计算速度也比 K-最近邻算法快得多, 从而成为目前使用最多的方法. 但是, 支持向量机学习的主要问题是低增量学习能力, 为了学习新模式必须重新训练; 而当得到新的数据模型时, 系统通常覆盖前面的模型; 与此同时, 在大多数情况下, 当训练样本量很大时, 需要消耗大量的时空资源^[6]. 因此, 用已知的概率分布替代具体的数据对支持向量机进行训练更合适.

本文提出了基于概率测度的支持向量机算法,

收稿日期: 2014-05-23; 修回日期: 2014-07-21

基金项目: 山东省科技发展计划资助项目(2014GGX101044)

并在此基础上提出了基于概率分布的静态手写数字识别方法. 实验表明: 本文方法能够在保证识别率的基础上大幅度提高运算速度.

2 经典的支持向量机和核理论

支持向量机是一种基于统计学理论的机器学习方法, 它将向量映射到高维空间, 在高维特征空间使用线性函数建立最大间隔超平面, 从而实现原始数据的分类, 其两分类问题算法如下:

对于独立同分布的训练样本集 $\{X_i, y_i\}_{i=1}^N$, 这里 $X_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$, N 是样本的个数, n 是每个训练样本向量的维数, 特征映射函数 $\phi(X_i): \mathbb{R}^n \rightarrow H$, 将训练样本集从输入空间 $\mathbb{R}^n$ 映射到高维特征空间 H , 它的决策过程只依赖于高维特征空间 H 中向量内积的运算, 即 $\langle \phi(X_i) \cdot \phi(X_j) \rangle$. 若存在核函数 $K(X_i, X_j)$ 满足 $K(X_i, X_j) = \langle \phi(X_i) \cdot \phi(X_j) \rangle$, 则它对应的分类决策函数为

$$f = \sum_{i=1}^N a_i y_i K(x_i, \cdot).$$

支持向量机在训练与决策过程中, 只需用核函数运算代替样本向量的内积运算, 没有必要知道 ϕ 的具体表达形式.

设 χ 为非空紧集, H 为函数族 F 的完备内积空间 (即希尔伯特空间), 且对于任意 $f \in F$ 有 $f: x \rightarrow \mathbb{R}$, 其中 $x \in \chi$. 如果对于所有 $x \in \chi$, 线性点函数映射 $f \rightarrow f(x)$ 存在且连续, 则 H 可称为一个再生核希尔伯特空间 (RKHS). 在 RKHS 中, $f(x)$ 可表示为一个内积: $f(x) = \langle f, \phi(x) \rangle_H$, 其中, $\phi: \chi \rightarrow H$ 为从 x 到 H 的特征空间映射, 且两个特征映射的内积称为核 (Kernel), $k(x, x') = \langle \phi(x), \phi(x') \rangle_H$.

3 优化的支持向量机

3.1 相关概念及问题描述

对于一个模式分类问题, 给定一非空紧集 χ, p 表示在测度空间 (χ, A) 中概率测度 P 的集合, A 是 χ 子集的一个 σ -代数. 我们的任务是: 当输入样本集 $\{(P_i, y_i)\}_{i=1}^m (P_i \in p, y_i \in Y)$ 时, 寻找能够最好地代表输入概率测度 P_i 与分类标签 y_i 关系的分类器 $h: p \rightarrow Y^{[7]}$. 为简单起见, 我们讨论两分类问题; 对于多分类问题, 可以使用 OAO (one against one) 或 OAA (one against all) 策略分解为多个两分类问题.

根据核学习理论, 我们将概率测度嵌入一个再生核希尔伯特空间 (RKHS), 从而 p 到 H 映射可表示为如下形式.

设 H 为函数 $f: \chi \rightarrow \mathbb{R}$ 的一个 RKHS, 根据 RKHS 的定义则必有再生核 $k: (\chi \times \chi \rightarrow \mathbb{R})$, 从而得到 p 到 H 均值映射

$$\mu: p \rightarrow H, P \mapsto \int_{\chi} k(x, \cdot) dP(x) \quad (1)$$

假设对所有 $x \in \chi, k(x, \cdot)$ 都是有限的, 如果 k 是特征核 (CK), 则 $p \rightarrow H$ 是内射的^[8]. 对于任意 P , 令 $\mu_P = \mu(P)$, 则有再生特征

$$E_P[f] = \langle \mu_P, f \rangle_H, \forall f \in H^{[9]}. \quad (2)$$

从式(2)可以看出: 均值嵌入 μ_P 作为与核函数 $K: (p \times p \rightarrow \mathbb{R})$ 相关联的特征映射, 可以定义 $K(P, Q) = \langle \mu_P, \mu_Q \rangle_H$. 由 $\sup_x \|k(x, \cdot)\|_H < \infty$, 得出

$$K(P, Q) = \iint_H \langle k(x, \cdot), k(z, \cdot) \rangle_H dP(x) dQ(z) \quad (3)$$

又由 H 的再生性得出:

$$K(P, Q) = \iint k(x, z) dP(x) dQ(z) \quad (4)$$

由式(3)和(4)得出 K 是关于 p 的正定核, 于是得到如下定义:

设训练样本为 $\{P_i, y_i\}_{i=1}^m$, 其中 $P_i \in p, y_i \in \mathbb{R}, \Omega: [0, +\infty) \rightarrow \mathbb{R}$ 为严格单调递增函数, 任一损失函数 $L: (p \times \mathbb{R}^2)^m \rightarrow \mathbb{R} \cup \{+\infty\}$, 正规风险函数定义为

$$L(P_1, y_1, E_{P_1}[f], \dots, P_m, y_m, E_{P_m}[f]) + \Omega(\|f\|_H) \quad (5)$$

使式(5)达到最小的 $f \in H$, 可表示为 $f = \sum_{i=1}^m a_i \mu_{P_i}$.

从上式可以看出, 关于概率测度的分类问题可以用嵌入 μ_p 的有限的线性组合函数来表式. a_i 为权重系数, 它通过嵌入 μ_p 来控制 f 中每个线性核的相对重要性.

3.2 核函数的确定

标准的 SVM 特征映射 $\phi(x)$ 一般是非线性的, 而 μ_p 是线性的. 所以在建模的不同阶段, 用两个核函数: 第一层嵌入核函数 k , 用来获得测度的向量表示; 第二层核函数 K 用来构建关于概率测度的非线性算法.

我们可以仿照定义非线性核 χ 的方法来定义非线性核 p , 将 $P \in p$ 的均值嵌入 μ_p 作为特征表示. 假设映射(1)是单射的, $\langle \cdot, \cdot \rangle_p$ 表示 p 的内积, 由线性得出 $\langle P, Q \rangle_p = \langle \mu_P, \mu_Q \rangle_H^{[10]}$. 关于 p 的非线性核定义为 $K(P, Q) = k(\mu_P, \mu_Q) = \langle \psi(\mu_P), \psi(\mu_Q) \rangle_{H_k}$, k 是正定核. 这样, 只要非线性核 p 完全依靠内积 $\langle \mu_P, \mu_Q \rangle_H$, 于是, 关于 χ 的标准非线性核就可以直接用来表示关于 p 的非线性核了. 这比针对确定概率测度而专门设计核函数简单灵活得多的. 并且, A .

Christmann 和 I. Steinwart 已经证明^[11]: 对于任意 $P, Q \in \mathcal{P}$, 高斯径向基核函数 $K(P, Q) = \exp\left(-\frac{\gamma}{2} \|\mu_P - \mu_Q\|_H^2\right)$ 是通用的, 其中 x 为一紧域, 映射 μ 是单射的。

在通常情况下, PM-SVM 能用扩展核函数 $K(P, Q) = E_{x \sim P, z \sim Q}[k(x, z)](E_{x \sim P}[\cdot])$ 指服从概率测度 P 的数学期望算子) 来处理解决 SVM 问题。当给定分布类和嵌入核函数 k 后, 由上式很容易计算出 $K(P, Q)$ 。例如: 当给定高斯分布类 $N(m, \Sigma)$, 嵌入核 k 为高斯径向基核函数 $\exp\left(-\frac{\gamma}{2} \|x - y\|^2\right)$ 后, 相对应的二级核函数 $K(P_i, P_j) = \langle \mu_{P_i}, \mu_{P_j} \rangle_H$ 为 $\exp\left(-\frac{1}{2} (m_i - m_j)^T (\Sigma_i + \Sigma_j + \gamma^{-1} I)^{-1} (m_i - m_j) / |\gamma \Sigma_i + \gamma \Sigma_j + I|^{\frac{1}{2}}\right)$ 。

4 实验与结论

实验中, 我们采用 MATLAB libsvm 工具箱(可从 <http://www.csie.ntu.edu.tw/~cjlin/libsvm/> 处下载), 在 MNIST 手写体数字库(可从 <http://yann.lecun.com/exdb/mnist/> 处下载)上进行实验。MNIST 是美国国家标准和技术研究所提供的一个专门用于手写体数字识别研究的数据库, 学术界许多实验研究基于此库。库里面包含了 60 000 个训练样本和 10 000 个测试样本, 每个样本为 28×28 像素的 bmp 图片^[12]。

对图像识别来说, 对输入图像进行不变性变换

不会影响对图像的识别结果。所谓的不变性变换是指对图像的缩放、平移、旋转等基本操作; 对训练样本施加不变性变换产生新的训练样本称为虚拟样本^[13]。本实验分为四组: 数字 1 和 8、3 和 4、3 和 8、6 和 9。对于每次训练, 我们从 MNIST 中, 对每个数字随机选取 100 个样本, 为了降低计算量, 用主成分分析法将维数降到 16, 对每个 16×16 的图像, 选用高斯分布 $N([1, 1], 0.1 \cdot I_2)$, $N([0, 0], 5 \cdot I_2)$ 和 $N(0, \pi)$ 作为先验分布, (s_x, s_y) 作为缩放参数, (t_x, t_y) 作为平移参数 (x, y 分别表示横纵坐标), θ 作为旋转角, 进行等价变换产生 10、20、30 个虚拟样本, 相应构建成为 2 000、4 000、6 000 个虚拟样本的训练集。

实验中, 我们对三种算法进行比较: 经典 SVM; 每一个分布选取相同数量样本的 SVM, (下称为 ASVM); 本文提出的基于测度 $\{(P_i, y_i)\}_{i=1}^m$ 的 PM-SVM 算法。对于 SVM 和 ASVM, 用高斯 RBF 核函数, PM-SVM 采用具有高斯 RBF 核函数的经验核做为基础核函数, 采用 10 CV 交叉验证法, 预先设定核参数搜索范围为: $C \in \{2^{-3}, 2^{-2}, 2^{-1}, \dots, 2^7\}$, $\gamma \in \{0.01, 0.1, 1\}$ 。表 1(a)、(b)、(c)。分别展示了本算法对缩放、平移、旋转变换后的图像的识别率, 在多数情况下, 当虚拟样本数量增大的时候, PM-SVM 比 SVM 和 ASVM 识别率更高一些; 表 2 中给出了以 PM-SVM 时间消耗为基准的相对时间消耗, 从中可以看出, 与 SVM 和 ASVM 相比, PM-SVM 算法的效率大幅提升。

表 1 识别准确率对比实验

(a) 缩放变换

算法	ASVM			SVM			PM-SVM		
虚拟样本数量	2 000	4 000	6 000	2 000	4 000	6 000	2 000	4 000	6 000
1 VS 8 / %	91.57	92.46	93.05	96.94	96.89	96.87	97.96	97.64	98.38
3 VS 4 / %	91.65	92.79	94.03	96.15	96.16	96.17	97.76	96.82	98.19
3 VS 8 / %	93.34	92.23	90.77	96.81	96.74	96.66	95.51	95.94	95.24
6 VS 9 / %	93.33	91.91	91.85	96.06	96.07	96.12	96.85	96.82	96.36

(b) 平移变换

算法	ASVM			SVM			PM-SVM		
虚拟样本数量	2 000	4 000	6 000	2 000	4 000	6 000	2 000	4 000	6 000
1 VS 8 / %	96.21	97.22	97.47	95.44	95.45	95.48	96.74	97.61	98.04
3 VS 4 / %	93.69	94.31	92.88	97.32	97.34	97.32	98.07	98.31	98.64
3 VS 8 / %	92.35	92.87	93.23	97.27	97.29	97.28	96.72	96.76	98.57
6 VS 9 / %	93.19	94.08	94.97	97.87	97.86	97.89	98.59	98.78	98.31

续表 1

(c) 旋转变换

算法	ASVM			SVM			PM-SVM		
虚拟样本数量	2 000	4 000	6 000	2 000	4 000	6 000	2 000	4 000	6 000
1VS 8 / %	96.92	93.50	96.54	92.74	92.76	92.80	95.34	96.13	94.61
3 VS 4 / %	95.63	95.56	91.23	93.92	93.87	93.93	96.58	94.57	94.88
3 VS 8 / %	92.36	92.76	92.65	96.53	96.53	96.53	93.64	96.53	97.84
6 VS 9 / %	94.15	95	94.48	99.25	99.23	99.23	98.63	98.63	99.36

表 2 ASVM 和 PM-SVM 识别时间

虚拟样本数量	2 000	4 000	6 000
SVM / %	465	1 291	2 089
ASVM / %	263	941	1 889
PM-SVM / %	1	1.48	1.16

5 结束语

本文针对静态手写数字识别问题,提出了一个基于概率测度的支持向量机算法(PM-SVM)。首先,将概率测度嵌入到再生核希尔伯特空间,得到优化的支持向量机模型;然后,针对该模型给出了关于概率测度的线性或非线性核函数,并根据实际情况选择适合的核函数,最后,实验证明了该算法在保证识别率的前提下,能大幅度提高计算效率,在总体性能上优于经典 SVM。

参考文献:

[1] S Impedovo, F M Manginin, D BarbuZZi. A novel prototype generation technique for handwriting digit recognition [J]. Pattern Recognition, 2014 (47): 1002-1010.

[2] 肖亚红. 基于 SVM 的静态手写签名识别方法研究[J]. 微电子学与计算机, 2011, 28(9): 135-138.

[3] 李琼, 陈利, 王维虎. 基于 SVM 的手写体数字快速识别方法研究 [J]. 计算机技术与发展, 2014, (2): 205-208.

[4] Mizukami Y, Yamaguchi K, Warrell J, et al. CUDA implementation of deformable pattern recognition and its application to MNIST handwritten digit database [C]// Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE, 2010: 2000-2004.

[5] Zhihong Man, Kevin Lee, Dianhui Wang, et al. An optimal weight learning machine for handwritten digit image recognition [J]. Signal Processing, 2013 (93): 1624-1638.

[6] X X Niu, C Y Suen. A novel hybrid CNN - SVM classifier for recognizing handwritten digits[J]. Pattern Recognition, 2012 (45): 1318-1325.

[7] 陶剑文, 王士同. 多核局部领域适应学习[J]. 软件学报, 2012, 23(9): 2297-2310.

[8] B K Sriperumbudur, A Gretton, K Fukumizu, et al. Hilbert space embeddings and metrics on probability measures[J]. Journal of Machine Learning Research, 2010(99): 1517-1561.

[9] Yang YH, Speed T. Design issues for cDNA microarray experiments [J]. Nat Rev Genet, 2002 (8): 579-588.

[10] A Berline, Thomas C Agnan. Reproducing kernel Hilbert spaces in probability and statistics [M]. USA: Kluwer Academic Publishers, 2004.

[11] A Christmann, I Steinwart. Universal kernels on non-standard input spaces[J]. Proceedings of Advances in Neural Information Processing Systems (NIPS), 2010 (3): 406-414.

[12] Ernst Kussul, Tatiana Baidyk. Improved method of handwritten digit recognition tested on MNIST database[J]. Image and Vision Computing, 2004 (22): 971-981.

[13] 魏兴国. 基于核方法的手写体数字识别研究[D]. 南京: 南京理工大学 2003.

作者简介:

方向女, (1978-), 硕士, 副教授. 研究方向为智能信息处理、网络安全. E-mail: fanxiang782@sina.com.