一种基于聚类核的半监督支持向量机分类方法*

李 涛, 汪西莉
(陕西师范大学 计算机科学学院, 西安 710062)

摘 要: 为了在标记样本数目有限时尽可能地提高支持向量机的分类精度,提出了一种基于聚类核的半监督支持向量机分类方法。该算法依据聚类假设,即属于同一类的样本点在聚类中被分为同一类的可能性较大的原则去对核函数进行构造。采用 K-均值聚类算法对已有的标记样本和所有的无标记样本进行多次聚类,根据最终的聚类结果去构造聚类核函数,从而更好地反映样本间的相似程度,然后将其用于支持向量机的训练和分类。理论分析和计算机仿真结果表明,该方法充分利用了无标记样本信息,提高了支持向量机的分类精度。

关键词: 聚类核; 聚类假设; 半监督支持向量机; 分类
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)01-0042-04
doi:10.3969/j.issn.1001-3695.2013.01.009

Semi-supervised SVM classification method based on cluster kernel

LI Tao, WANG Xi-li
(*School of Computer Science, Shaanxi Normal University, Xi'an* 710062, *China*)

Abstract: In order to improve the classification accuracy of support vector machine when limited the labeled samples, this paper proposed a semi-supervised support vector machine classification method. It constructed the kernel function according to the cluster assumption. The method used both the labeled and the unlabeled samples to construct the kernel function by the K-means algorithm. The similarity between the samples could be represented better by such kernel function. Then it used in SVM to train and obtain the classification results. Theoretical analysis and computer simulation results show that the algorithm can effectively use a large number of unlabeled samples, and can improve the classification accuracy.

Key words: cluster kernel; cluster assumption; semi-supervised support vector machine; classification

0 引言

在一些实际应用中,收集到的观测样本大多是没有类别标记的,对样本进行类别标记往往需要耗费大量的人力物力。如果只使用少量的有标记样本进行训练,利用其所训练出的学习系统往往很难具有强的泛化能力,在这种情况下需要解决的是如何利用已标记样本和无标记样本所蕴涵的信息来对无标记样本进行合理的标记。因而如何将无标记样本引入学习中,是机器学习的研究热点,半监督学习的目的则是将少量标记样本与大量无标记样本相结合以提高学习器的泛化能力。

支持向量机能够较好地解决高维数据的非线性分类问题,因而被广泛地使用[1]。标准的支持向量机是基于监督学习的,Joachims[2]将半监督学习引入到支持向量机中,形成了直推式支持向量机 TSVM。其可以预测潜在的未知数据,因而 TSVM 在通常情况下也被认为是 Semi-Supervised SVM (S^3VM)。由于 S^3VM 同时利用已标记和未标记样本去最大化分类间隔,从而使得其目标函数是非凸的[3]。目前,S^3VM 的研究主要集中在非凸目标函数的优化上。Chapelle 等人[4]利用梯度下降法来解决非凸优化问题,提出了梯度下降 S^3VM。随后 Chapelle 等人[5]又通过连续优化技术对半监督支持向量机的非凸优化问题进行求解,提出 CS^3VM。Collobert 等人[6]通过凹凸过程来解决非凸优化问题,提出了 CCCP S^3VM。

这些算法主要解决的是 S^3VM 目标函数的非凸优化问题,需要反复迭代运算,计算复杂度较高,难以应用于大规模数据的分类[7]。其半监督思想主要体现为:在标准支持向量机的目标函数中加入了无标记样本的损失项,使用混合样本进行学习,经过求解使得目标函数达到极小值的最优决策超平面来得到一个泛化性能更好的分类器。最终达到通过利用无标记样本的信息来调整决策边界,从而得到一个既通过数据相对稀疏的区域又尽可能正确划分有标记样本的超平面。

而根据聚类假设,如果高密度区域的两个点可以通过区域内某条路径相连接,那么这两点拥有相同标记的可能性就比较大[8]。在此基础上,Chapell 等人[9]提出了构造聚类核的整体框架,其具体过程是根据已有的标记样本和无标记样本去构造核矩阵,通过使用不同的转换函数去改变核矩阵经过特征分解后的特征值来得到不同的核,如随机游走核和谱聚类核。其中随机游走核依据的是马尔可夫随机游走的思想,将样本点看成是一个全连通图上的节点,将两节点之间的距离转换为随机游走过程中两状态之间的转移概率,便可以得到一个 $(m+n)\times(m+n)$ 的转移矩阵,其中 $m+n$ 是样本点的数目。想象每个节点上都存在一个可以运动的粒子,它们按转移概率随机游走到其他节点。这样某个粒子 x_i 在 t 步转移之后的概率分布也就是 t 阶转移矩阵的第 i 行 p_i^t,从而也就知道某个起始点在 x_i 的粒子经过 t 步转移之后到达节点 x_j 的概率为 p_{ij}^t。由于转移

收稿日期: 2012-06-08; **修回日期**: 2012-07-09 **基金项目**: 国家自然科学基金资助项目(41171338)

作者简介:李涛(1984-),男,陕西渭南人,硕士研究生,主要研究方向为智能信息处理、模式识别(litaostudy@126.com);汪西莉(1969-),女,教授,博导,博士,主要研究方向为智能信息处理、模式识别、图像处理.

概率矩阵可以识别出具有高的类内转移概率和低的类间转移概率的子集,从而可以避免初始分布的不确定性,进而更好地反映数据之间的真实关系。因而在随机游走核中,根据样本点之间的近邻关系建立相似度矩阵,计算 t 步转移概率矩阵,并将其用于核矩阵的计算中。其中 t 主要用于控制在新的特征空间中样本之间分布的稠密程度。通过 t 的设置,可以使得应属于同一类的样本点之间的分布更加密集。谱聚类核依据的是谱聚类的思想,即通过对样本之间的相似度矩阵进行谱分解,对样本点在特征空间中进行重新表示,使得位于同一聚类或高密度区域中的样本点在新的空间中分布得更紧凑。

而随机游走核和谱聚类核所存在的问题是必须对一个 $(m+n)\times(m+n)$ 的矩阵进行对角化,其中 m 和 n 分别是标记样本和无标记样的数目,其对角化的时间复杂度为 $O(m+n)^3$。Jason Weston、Christina Leslie 等人提出了通过多次 K-均值聚类去构造 bagged 聚类核[9],该方法的思想是通过 K-均值聚类算法估计两两样本位于同一聚类的概率,进而得到样本间的相似度信息,其时间复杂度为 $O(rk(m+n)d)$。其中:d 是样本数据的维数,r 是 K-均值聚类算法的运行次数。

根据上述分析,本文提出了一种基于 bagged 聚类核的半监督支持向量机分类方法。该方法依据聚类假设的思想,对已有的标记样本和所有的无标记样本采用多次 K-均值聚类运算去构造 bagged 聚类核,然后对 bagged 聚类核和径向基核进行求和,并最终用于支持向量机的训练和分类中。该方法的算法复杂度为线性阶,较随机游走核和谱聚类核的复杂度低,训练中采用了所有的标记样本和无标记样本,从而可以更准确地反映出样本空间的整体分布。本文方法的创新之处为:在构造半监督支持向量机的过程中,只需要构造一个 $m\times(m+n)$ 的矩阵,而不需要去对角化一个 $(m+n)\times(m+n)$ 的矩阵。与基于随机游走核和谱聚类核的半监督支持向量机相比,在构造聚类核的过程中,空间复杂度由 $O(m+n)^2$ 变为 $O(m\times(m+n))$,时间复杂度由 $O(m+n)^3$ 变为 $O(rk(m+n)d)$,降低了构造聚类核的空间复杂度和时间复杂度。同时,在构造聚类核过程中,可以使用所有无标记样本信息,从而尽可能地提高分类准确率。通过数据集实验和图像实验,证实了本文算法的有效性。

1　支持向量机分类

支持向量机是一种基于统计学习理论的机器学习方法,它根据有限的样本信息在模型的复杂性与学习能力之间寻求最佳折中,以期获得最好的推广能力。它通过将低维空间的非线性问题映射到高维空间,进而可以用简单的线性分类技术对样本进行分类,并且具有很强的泛化能力。通过核函数代替特征空间中的点积运算,因而不必具体知道映射函数的形式。

支持向量机分类方法的数学描述如下:对于给定的一组训练样本

$$\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\},x_i\in R^N,y_i\in\{-1,+1\}$$

可以通过一个非线性映射 $\Phi(\cdot)$,将训练样本映射到一个高维空间。$\Phi:R^N\to H$。对于非线性情况,支持向量机最优分类面的求解问题则可以表示为

$$\min\{\frac{1}{2}\|w\|^2+C\sum_i\xi_i\}\tag{1}$$

$$y_i(\langle\varphi(x_i),w\rangle+b)\geqslant 1-\xi_i\quad\forall i=1,2,\cdots,n\tag{2}$$

$$\xi_i\geqslant 0\quad\forall i=1,2,\cdots,n\tag{3}$$

上述式(1)约束于式(2)。其中,$2/\|w\|$ 为分类间隔;b 为常数;C 是惩罚系数;ξ_i 是松弛因子。

通常求解式(1)时采用求取其拉格朗日对偶问题的方法。

$$\max_{\alpha_i}\{\sum_i\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_jy_iy_jK(x_i,x_j)\}$$
$$0\leqslant\alpha_i\leqslant C,\sum_i\alpha_iy_i=0\quad i=1,2,\cdots,n\tag{4}$$

其中:$K(x_i,x_j)$ 为核函数。

对于任意一个测试向量 x,决策函数为

$$f(x)=\mathrm{sgn}\ (\sum_{i=1}^{n}y_i\alpha_iK(x_i,x)+b)\tag{5}$$

其中:α_i 为拉格朗日乘子,支持向量为训练样本中非零 α_i 所对应的 x_i。

在支持向量机中常用的核函数有线性核函数 $K(x,z)=\langle x,z\rangle$,多项式核函数 $K(x,z)=(\langle x,z\rangle+1)^q$,$q$ 为整数。径向基核函数 $K(x,z)=\exp(-\|x-z\|^2/2\sigma^2)$,$\sigma\in R^+$。

2　基于 bagged 聚类核的支持向量机分类

2.1　Bagged 聚类核

为了简化问题,依据聚类假设思想:位于同一聚类或高密度区域的样本应该有相同的标记,若两个样本位于同一聚类中,则这两个样本之间的距离就越小,相似度就越大;若两个样本位于不同的聚类中,则这两个样本之间的距离就越大,相似度就越小。因而可以通过 K-均值聚类算法对两个样本之间的相似程度作出一个概率估计,构造 bagged 聚类核[9]。Bagged 聚类核的思想则是通过采用 K-均值聚类算法,对两两样本位于同一聚类的概率作出一个估计,使得位于同一聚类中或高密度区域中的样本间的相似度越大,而位于不同聚类中的样本间的相似度越小。

Bagged 聚类核的构造过程如下:

a)运行 N 次 K-均值聚类运算,对于样本点,每次的聚类结果为 $c_j(x_j)$,其中 j 为 N 次 K-均值迭代运算,$j=1,2,3,\cdots,N$。其中 x 是所有的标记样本和无标记样本。每次的 K-均值聚类中心都是随机产生的。

b)建立 bagged 聚类核

$$K_{\mathrm{bag}}(x,x')=\frac{\sum_j[c_j(x)=c_j(x')]}{N}$$

其中,在第 j 次迭代过程中,若 x 和 x' 在同一类中,则 $c_j(x)=c_j(x')$ 返回值为 1,否则返回值为 0。$K_{\mathrm{bag}}(x,x')$ 是一个满足 Mercer 定理的核,它等价于在 N_k 维空间中,特征函数 $\varphi(x_i)=\langle[c_j(x_i)=q]:j=1,\cdots,N;q=1,\cdots,k]$ 之间的内积[10]。

通过分析上述 bagged 聚类核的表达式可看出,该核主要由 K-均值聚类和迭代次数 N 决定,其中聚类数目可以反映出数据的分布程度,当 k 值越大,则对于数据集得到的聚类块数越多,从而使得更有可能位于同一聚类中的数据间的相似程度更大,而位于不同聚类中的数据间的相似程度越小。迭代次数 N 则是用于在运行 K-均值聚类算法后,两两样本位于同一聚类次数的一个概率平均估计。在本文中 $N=50$。对于 k 值则在下文的实验中,选取较高的分类精度所对应的 k 值。

2.2　基于 bagged 聚类核的半监督支持向量机算法

本文采用标记样本和所有的无标记样本信息来构造支持

向量机的核函数,使得在融入了无标记样本信息后,标记样本之间的相似度能够增强核函数的适应性。基于聚类核的半监督支持向量机分类方法首先对标记样本和所有无标记样本进行聚类构建一个 bagged 聚类核,随后去和径向基核进行求和运算。以下为基于 bagged 聚类核的分类算法步骤:

a)设标记样本的数目为 m,无标记样本的数目为 n。构建一个 $m\times(m+n)$ 的零矩阵。其中非标记样本是所有的测试样本。

b)使用 RBF 核函数,计算标记样本集的核矩阵 K_{train},则由标记样本所组成的核矩阵 K_{train} 为 $m\times m$ 的矩阵。使用 RBF 核函数计算标记样本和无标记样本所组成的测试核矩阵,则测试核矩阵 K_{test} 为 $m\times n$ 的矩阵。

c)样本集合 X 为由标记样本和无标记样本所组成的集合。对样本集合 X 运行 N 次 K-均值聚类算法。

d)根据聚类结果建立 bagged kernel:

$$K_{\text{bag}}(x_i,x_j)=\frac{[C(x_i)=C(x_j)]}{N}$$

其中:$i=\{1,2,\cdots,m\}$,$j=\{1,2,\cdots,m+n\}$。如果 x_i 和 x_j 在同一类中,则$[C(x_i)=C(x_j)]$返回值 1,否则返回值 0,得到的 K_{bag} 为 $m\times(m+n)$ 的矩阵。

e)根据矩阵 K_{bag} 得到由前 m 列组成的 $m\times m$ 的矩阵 K'_{bag} 和由后 n 列组成的 $m\times n$ 的矩阵 K_{testbag}。

f)对 K_{train} 和 K'_{bag} 进行求和,α 为权重因子,得到训练核矩阵 $K(x_i,x_j)$。其中,K'_{bag} 是由步骤 d)中得到的 K_{bag} 对应标记样本所构成的矩阵。$i=\{1,2,\cdots,m\}$,$j=\{1,2,\cdots,m\}$。

g)采用核矩阵 $K(x_i,x_j)$ 去训练支持向量机。

h)对 K_{test} 和 K_{testbag} 进行求和,α 为权重因子,得到测试核矩阵 $K(x_i,x_j)'$。

$$K(x_i,x_j)'=\alpha K_{\text{testbag}}(x_i,x_j)+(1-\alpha)K_{\text{test}}(x_i,x_j)$$

其中:$i=\{1,2,\cdots,m\}$,$j=\{1,2,\cdots,n\}$。

i)采用步骤 g)所得的支持向量机分类器和步骤 h)所得到的测试核矩阵,完成对测试样本的分类。

3 实验与分析

本文分别对数据集、人工合成图像以及自然图像进行了相关实验。在所有的实验中,支持向量机的参数调整采用的是网格搜索方法。范围分别是:高斯核函数参数 $\sigma=\{10^{-3},\cdots,10^{0}\}$,惩罚系数 $C=\{10^{-3},\cdots,10^{3}\}$。参数 a 的取值为 0.5。实验 1~3 为数据集的分类实验。实验 1 所采用的数据集为人工生成的 toy 数据集。Toy 数据集是一个双月型数据集[11],上半月和下半月各表示一类。实验 2 所采用的数据集为 UCI 数据集 sonar,共有两类。实验 3 所采用的数据集为 UCI 数据集 Iris,共有三类。上述数据集和后续图像实验中的测试样本与各自的无标记样本一致。实验 4~7 为人工合成图像分类。实验 8~11 则为自然图像分类,图像 1~4 尺寸大小为 256×256 像素。图像 5~6 均来源于 BSD 图像库[12],尺寸大小为 481×321 像素。图像 7~8 来源于 GrabCut 图像库[13],尺寸大小为 640×480 像素。

图 1~11 为 K-均值聚类算法的聚类数目对分类结果的影响。表 1 为本文方法的参数选择,表 2 为标准 SVM 的参数选择,表 3 给出了内存 1 GB、MATLAB 7.0 环境下本文方法(BAG-SVM)及标准 SVM 方法运行 20 次后的分类精度和平均运行时间。

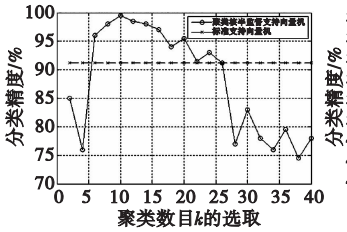


图1 Toy数据集的k值选择与分类精度的关系

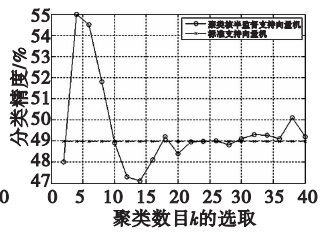


图2 Sonar数据集的k值选择与分类精度的关系

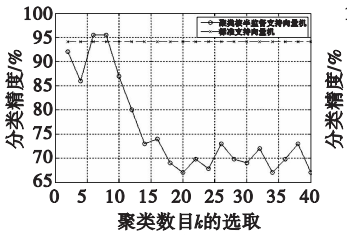


图3 Iris数据集的k值选择与分类精度的关系

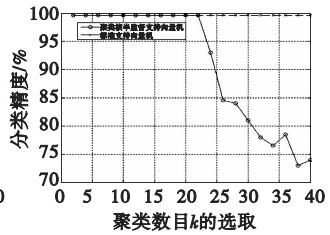


图4 图像1的k值选择与分类精度的关系

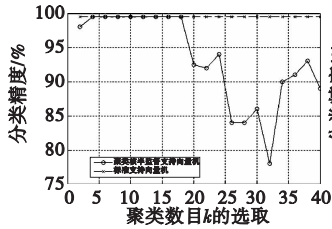


图5 图像2的k值选择与分类精度的关系

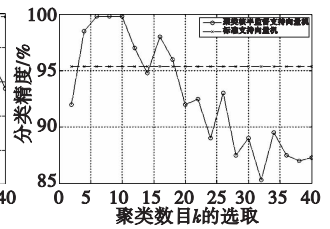


图6 图像3的k值选择与分类精度的关系

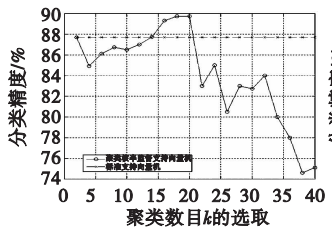


图7 图像4的k值选择与分类精度的关系

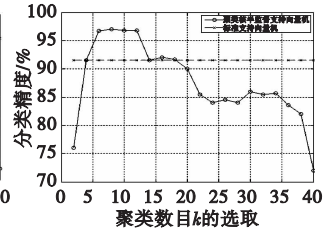


图8 图像5的k值选择与分类精度的关系

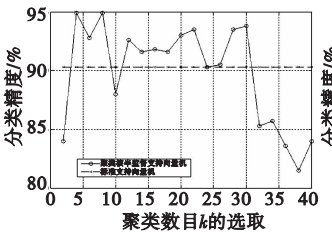


图9 图像6的k值选择与分类精度的关系

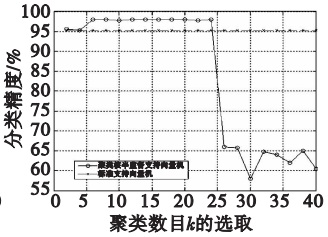


图10 图像7的k值选择与分类精度的关系

表 1 本文方法的参数选择和使用的样本数目

实验	聚类数目 k	核参数 σ	惩罚系数 C	每类标记样本	无标记样本数
1	10	0.003 5	1 00 0	5	182
2	4	0.001 0	831.251 5	5	198
3	8	0.001 0	954.845 5	5	135
4	2	0.005 0	1 000	10	65 536
5	3	0.001 0	1 000	10	65 536
6	10	0.001 0	1 000	10	65 536
7	18	0.012 1	1 000	10	65 536
8	12	0.001 0	1 000	10	154 401
9	8	0.001 6	757.876 9	10	154 401
10	8	0.012 1	954.845 5	10	307 200
11	26	0.001 0	723.655 3	10	307 200

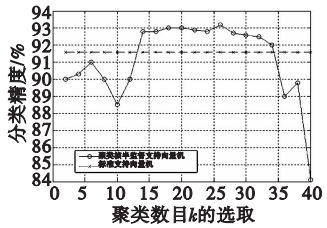


图11　图像8的k值选择与分类精度的关系

表 2　标准支持向量机的参数选择和使用的样本数目

实验	核参数 σ	惩罚系数 C	每类标记样本	实验	核参数 σ	惩罚系数 C	每类标记样本
1	0.003 5	1 000	5	7	0.012 1	831.251 5	10
2	0.001 0	757.876 9	5	8	0.001 0	954.845 5	10
3	0.001 0	870.561 2	5	9	0.001 6	601.539 6	10
4	0.005 0	1 000	10	10	0.012 1	870.561 2	10
5	0.001 0	1 000	10	11	0.001 0	629.986 3	10
6	0.001 0	954.845 5	10				

表 3　本文方法和标准 SVM 的分类结果

实验	分类方法	平均准确率/%	最大准确率/%	平均运行时间/s
1	本文方法	98.82	99.45	0.260
	标准 SVM	91.21	91.21	0.010
2	本文方法	53.61	56.57	0.370
	标准 SVM	48.99	48.99	0.010
3	本文方法	95.78	96.30	0.170
	标准 SVM	94.07	94.07	0.005
4	本文方法	99.75	99.75	6.103
	标准 SVM	99.63	99.63	0.647
5	本文方法	99.51	99.51	13.160
	标准 SVM	99.45	99.45	0.790
6	本文方法	99.74	99.74	96.764
	标准 SVM	95.36	95.36	1.108
7	本文方法	89.72	90.42	340.292
	标准 SVM	87.67	87.67	1.240
8	本文方法	98.16	98.18	323.81
	标准 SVM	91.53	91.53	1.610
9	本文方法	94.02	94.92	463.310
	标准 SVM	90.30	90.30	5.220
10	本文方法	96.99	98.65	512.410
	标准 SVM	95.25	95.25	2.730
11	本文方法	93.28	93.34	4 353.030
	标准 SVM	91.59	91.59	12.100

图 12 为合成图像分类结果,图 13 为自然图像分类结果。

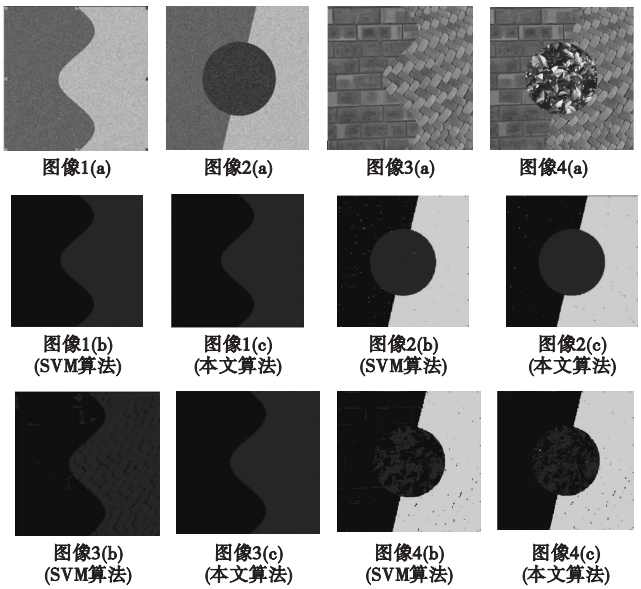


图12　合成图像分类结果

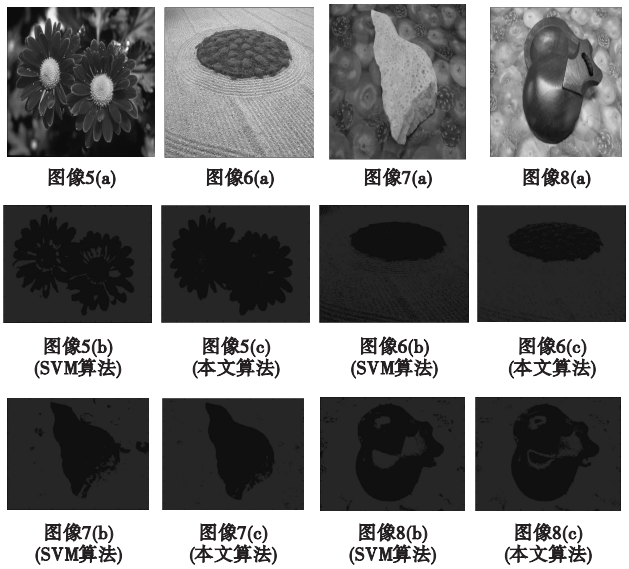


图13　自然图像分类结果

通过观察以上数据集和图像分类实验中 K-均值聚类的聚类数目与分类精度之间的曲线图(图 1 ~ 11),可以得出以下结论:对于数据集,当 k 值大于样本集所固有的类别数目时,采用本文方法较标准 SVM 的分类精度有所提高。对于图像分类,若图像较为平滑时,当 k 值等于图像所固有的类别数目时,采用标准 SVM 方法的分类精度与本文方法相当;若图像不平滑且较为复杂时,当 k 值大于图像所固有的类别数目时,采用本文方法可以有效地提高图像分类精度。从图像的分类结果中可以看出,对于平滑图像 1 ~ 2,两种方法的分类结果趋于一致;而对于复杂纹理图像 3 ~ 4、自然图像 5 ~ 8,采用本文方法得到的分类结果更符合区域一致性,优于标准 SVM 的分类结果。

图 14 给出了采用 bagged 核的直观解释,其中图 14(a)给出了待分数据的分布情况和理想的核矩阵,(b) ~ (d)给出了采用 K-均值聚类算法构造 bagged kernel 时不同的聚类数目得到的 bagged 核。当 k 值的选取与数据集的原始类别数目一致时,采用由标记样本和无标记样本组成的半监督核,使得原本并不在同一类中的样本之间相似度有增大的趋势。相比于仅由标记样本组成的监督核,采用由标记样本和无标记样本组成的半监督核并没有达到促使不同类样本之间的相似度变小的效果,从而 BAG-SVM 的分类精度并不优于标准 SVM 的分类精度。而当 k 值的选取大于数据集的原始类别数目时,采用由标记样本和无标记样本组成的半监督核相比于仅由标记样本组成的监督核,使得位于同一类别中的样本间的相似度更大,而不在同一类别中的样本间的相似度相应较小,进而更准确地反映了样本之间的相似程度,使得 BAG-SVM 的分类精度优于 SVM 的分类精度。

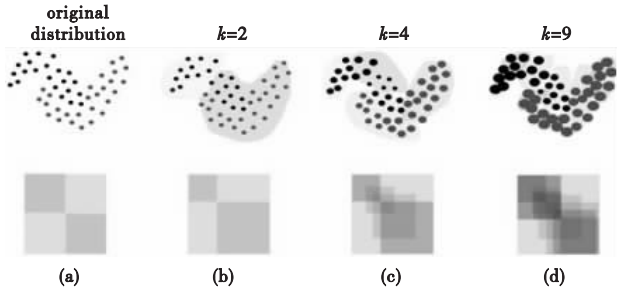
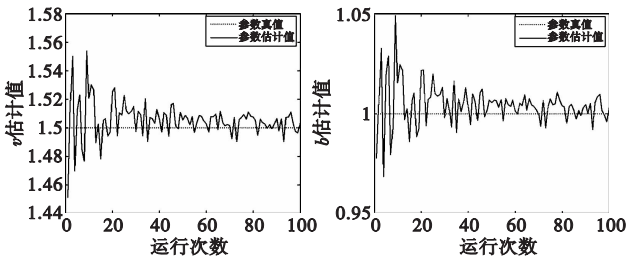


图14　Bagged核的直观解释

实验表明,在标记样本数目非常有限的情况下,使用本文方法可以提高 SVM 的分类精度。

(下转第 48 页)



(a)形状参数v的估计值 (b)尺度参数b的估计值

图3 对数累积量估计的精度随运行次数的变化关系（v=1.5，b=1，对于不同的运行次数，样本数量均为10 000）

表1 对数累积量估计的 Monte Carlo 仿真结果（$b=1$，独立运行100次，每次的样本数量均为10000）

参数真值	$\hat{v}$	$\hat{b}$
$v=0.5$	0.500 2 (0.008 6)	1.000 9 (0.038 1)
$v=1$	1.004 7 (0.024 2)	1.006 1 (0.033 8)
$v=1.5$	1.502 5 (0.053 9)	1.002 3 (0.047 5)
$v=2$	1.997 1 (0.102 2)	0.997 7 (0.063 4)
$v=5$	5.000 5 (0.649 6)	1.000 4 (0.138 2)

表2 对数累积量估计的 Monte Carlo 仿真结果（$v=1.5$，独立运行100次，每次的样本数量均为10000）

参数真值	$\hat{v}$	$\hat{b}$
$b=0.5$	1.496 8 (0.046 2)	0.499 0 (0.019 1)
$b=1$	1.501 1 (0.053 6)	0.999 9 (0.044 7)
$b=1.5$	1.501 3 (0.060 3)	1.503 8 (0.079 0)
$b=2$	1.501 6 (0.054 5)	2.003 4 (0.093 2)
$b=5$	1.502 5 (0.051 3)	5.004 1 (0.231 0)

5 结束语

为了高效地估计出 K 分布的参数，本文提出了基于第二类统计量的对数累积量估计方法。首先对 K 分布的概率密度函数进行 Mellin 变换，从而获得 K 分布的第二类第一特征函数；然后对第二类第一特征函数进行对数变换，从而获得 K 分布的第二类第二特征函数；最后对第二类第二特征函数求前两阶导数，进而获得 K 分布的对数累积量估计式。与传统的最大似然估计方法相比，对数累积量估计具有简洁的解析表达式，易于计算。Monte Carlo 仿真表明，对数累积量估计可以获得较高的估计精度。因此，基于第二类统计量的对数累积量估计是 K 分布参数估计的高效方法。作为强有力的数学工具，第二类统计量同样可以用于其他常见分布（如对数—正态分布）的参数估计中，这是下一步的研究内容。

参考文献：

[1] 张红，王超，张波，等. 高分辨率 SAR 图像目标识别[M]. 北京：科学出版社，2009.

[2] SKOLNIK M I. 雷达系统导论[M]. 3 版. 左群声，徐国良，马林，等译. 北京：电子工业出版社，2006.

[3] WARD K D，TOUGH R J A，WATTS S. Sea clutter：scattering，the K distribution and radar performance[M]. London：The Institution of Engineering Technology，2006.

[4] 皮亦鸣，杨建宇，付毓生，等. 合成孔径雷达成像原理[M]. 成都：电子科技大学出版社，2007.

[5] PAPOULIS A，PILLAI S U. 概率、随机变量与随机过程[M]. 保铮，冯大政，水鹏朗，译. 西安：西安交通大学出版社，2004.

[6] 孙增国，韩崇昭. 基于拖尾分布的高分辨率合成孔径雷达图像建模[J]. 物理学报，2010，59(2)：998-1008.

[7] ACHIM A，KURUOGLU E E，ZERUBIA J. SAR image filtering based on the heavy-tailed Rayleigh model[J]. IEEE Trans on Image Processing，2006，15(9)：2686-2693.

[8] NICOLAS J M. Alpha-stable positive distributions：a new approach based on second kind statistics[C]//Proc of European Signal Processing Conference. Toulouse：EURASIP，2002：197-200.

[9] TISON C，NICOLAS J M，TUPIN F，*et al*. A new statistical model for Markovian classification of urban areas in high-resolution SAR images[J]. IEEE Trans on Geoscience and Remote Sensing，2004，42(10)：2046-2057.

[10] GRADSHTEYN I S，RYZHIK I M. Table of integrals，series，and products[M]. 北京：世界图书出版公司，2007.

[11] PRESS W H，TEUKOLSKY S A，VETTERLING W T，*et al*. Numerical recipes in C[M]. Cambridge：Cambridge University Press，1994.

（上接第45页）

4 结束语

本文依据聚类假设的思想，提出了一种基于聚类核的半监督支持向量机分类方法。该算法根据高密度区域的样本在聚类中被分在同一类中的特性，采用 K-均值聚类算法对样本集进行多次聚类的方法构造聚类核，将无标记样本的信息融入分类器的训练过程中。该方法能够有效利用无标记样本信息，而且实验结果也证实了无标记样本的加入可以提高分类的精度。但本方法涉及到支持向量机中参数的选择、K-均值聚类算法的有效性，并且随着数据量的增大，运行时间也会随之增加。因此，如何更有效地进行参数选择和 K-均值聚类，同时降低时间复杂度仍然有待研究。

参考文献：

[1] SCHOLKOPF B，SMOLA A J. Learning with kernel：support vector machines，regularization，optimization and beyond[M]. Cambridge，MA：MIT Press，2002：190-195.

[2] JOACHIMS T. Transductive inference for text classification using support vector machines[C]//Proc of the 16th International Conference on Machine Learning. San Francisco：Morgan Kaufmann Publishers，1999：200-209.

[3] CHAPELLE O，SINDHWANI V，KEERTHI S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research，2008，9(2)：203-233.

[4] CHAPELLE O，ZIEN A. Semi-supervised classification by low density separation[C]//Proc of the 10th International Workshop on Artificial Intelligence and Statistics. San Francisco：Morgan Kaufmann Publishers，2005：57-64.

[5] CHAPELLE O，CHI M，ZIEN A. A continuation method for semi-supervised SVMs[C]//Proc of the 23rd International Conference on Machine Learning. New York：ACM Press，2006：185-192.

[6] COLLOBERT R，SINZ F. Large scale transductive SVMs[J]. Journal of Machine Learning Research，2006，7(8)：1687-1712.

[7] ZHU Xiao-jin. Semi-supervised learning literature survey TR1530[R]. Wisconsin：Dept. of Computer Sciences，University of Wisconsin，2008：13-16.

[8] SERGIOS T，KONSTANTIONS K. Pattern recognition[M]. [S. l.]：Academic Press，2008：577-582.

[9] CHAPELLE O，WESTON J，SCHOLKOPF B. Cluster kernels for semi-supervised learning[C]//Proc of the 16th Annual Conference on Neural Information Processing Systems. Massachusetts：MIT Press，2003：321-328.

[10] CHAPELLE O，SCHÖLKOPF B，ZIEN A. Semi-supervised learning[M]. Massachusetts：MITZIEN Press，2006：348-351.

[11] TUIA D，CAMPS-VALLS G. Urban image classification with semisupervised multiscale cluster kernels[J]. IEEE Journal of Selected topics in Applied Earth Observations and Remote Sensing，2011，4(1)：65-74.

[12] The berkeley segmentation dataset and benchmark[EB/OL][2011-02-20]. http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/.

[13] GULSHAN V，ROTHER C，CRIMINISI A，*et al*. Geodesic star convexity for interactive image segmentation[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2010：3129-3136.