一种基于半监督集成SVM的土地覆盖分类模型

刘 颖[1,2,3] 张 柏[2] 王爱莲[1] 桑 娟[1] 何咏梅[1]
(吉林财经大学管理科学与信息工程学院 长春 130117)[1]
(中国科学院东北地理与农业生态研究所 长春 130102)[2] (中国科学院研究生院 北京 100049)[3]

摘 要 目前,支持向量机技术(SVM)在遥感信息获取中普遍受到参数选择不准确和小样本问题的制约。针对这些问题,提出一种新的半监督集成SVM(EPS3VM)分类模型。模型一方面利用自适应变异粒子群优化算法对SVM参数寻优以提高基分类器精度(PSVM);另一方面采用自训练算法(Self-training),充分利用大量廉价的未标记样本产生性能差异的半监督分类器个体(PS3VM),其中,在未标记样本标注过程中,引入模糊聚类算法(Gustafson-kessel)来控制错误类别的输入,最后对个体分类器采用加权集成策略,以进一步提高分类模型的泛化能力。为了测试其性能,应用该模型进行多光谱遥感影像的土地覆盖分类实验,并与PSVM、PS3VM进行对比,分类精度从PSVM的88.48%提高到96.88%,Kappa系数由0.8546提高到0.9606。结果表明,EPS3VM在克服传统SVM参数选择不准确的同时,有效地应对了小样本问题,分类性能更优。

关键词 支持向量机,半监督学习,集成学习,Gustafson-kessel模糊聚类,土地覆盖,分类
中图法分类号 TP391.4 **文献标识码** A

Ensemble Model with Semisupervised SVM for Remote Sensing Land Cover Classification

LIU Ying[1,2,3] ZHANG Bai[2] WANG Ai-lian[1] SANG Juan[1] HE Yong-mei[1]
(College of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117, China)[1]
(Northeast Institute of Geography and Agroecology Chinese Academy of Sciences, Changchun 130102, China)[2]
(Graduate University of Chinese Academy of Sciences, Beijing 100049, China)[3]

Abstract Nowadays, most SVM-based remote sensing classification methods are challenged by incorrectly selecting parameters values and the small sample problems. This paper proposed a novel ensemble model with semisupervised SVM (EPS3VM) to address the problem of remote sensing images classification. The key characteristics of this approach are to 1) self-adaptive mutation particle swarm optimizer is introduced to improve the generalization performance of the SVM classifier (PSVM), 2) self-training semisupervised learning method that leverages large amounts of relatively inexpensive unlabeled data is presented to produce a number of semisupervised classifiers (PS3VM). Then by the weighted voting method, these classifiers are combined so as to improve the generalization ability of the classification model. In order to reduce the impact of this issue by incorrect labels, Gustafson-kessel fuzzy clustering algorithm (GKclust) is used for selecting the useful points from the unlabeled set. The effectiveness of the proposed classification approach is demonstrated for identifying different land cover regions in multispectral remote sensing imagery. In particular, the performance of the EPS3VM is compared with PSVM and PS3VM in terms of classification accuracy and kappa coefficient. On an average, the EPS3VM model yields an overall accuracy of 96.88% against 88.48% for PSVM and outperformed PS3VM in terms of overall accuracy (by about 5%). The obtained results clearly confirm the effectiveness and robustness of the EPS3VM approach to the remote sensing land cover classification.

Keywords Support vector machines, Semisupervised learning, Ensemble learning, Gustafson-kessel fuzzy clustering, Land cover, Classification

1 引言

土地利用是指人类施加于地表的活动,是基于人类活动的影响程度及活动目的进行划分的[1]。土地利用分类根据土地类型在遥感影像数据中表现出的光谱特征差异,进行土地覆盖归属和识别。目前许多机器学习算法被应用于多光谱土

到稿日期:2012-09-19 返修日期:2012-12-22 本文受国家重点基础研究发展计划(973计划)课题(2009CB421103),中国科学院重点部署项目课题(KZZD-EW-08-02),吉林省科技发展计划项目(20130522177JH)资助。

刘 颖(1979-),女,博士生,主要研究方向为模式识别、计算智能、遥感图像处理等,E-mail:lyaihua1995@163.com;张 柏(1962-),男,研究员,博士生导师,主要研究方向为遥感应用研究、智能模型构建等,E-mail:zhangbai@neigae.ac.cn(通信作者);王爱莲(1967-),女,副教授,主要研究方向为模式识别、数据库技术等;桑 娟(1964-),女,讲师,主要研究方向为人工智能等;何咏梅(1968-),女,副教授,主要研究方向为智能计算、计算机网络等。

地覆盖分类领域[2]，其中支持向量机技术(Support Vector Machines，SVM)由于能较好地解决高维特征、非线性、过学习，且具有局部极小等优点，成为土地覆盖信息获取技术新的研究热点，广泛应用于森林类型识别[3,4]、农业作物监测[5]、道路信息提取[6]、图像分割[7]等领域。

SVM尽管在遥感信息获取中取得了很好的效果[8]，但仍存在有待改进和完善之处，主要表现在以下两方面，1)参数选择问题：分类参数的选择没有特别好的办法，应用时不容易找到最优分类参数；2)小样本问题：当训练样本集远远小于测试样本集时，SVM即便具有较强的泛化性，也难以给出令人满意的结果。针对上述问题，诸多专家学者开展一系列卓有成效的工作，并取得丰硕的研究成果。

对于分类参数选择问题，常用的方法是网格法，然而，这种方法消耗大量时间，结果不是很理想[9]。很多学者利用智能优化算法对SVM参数进行优化，如采用遗传算法(Genetic Algorithms，GA)实现SVM特征子集的选择与参数同步优化[10]，用粒子群优化算法(Particle Swarm Optimization，PSO)构建SVM分类器优化模型[11]。在寻优过程中相比GA方法，PSO没有GA算法的选择、交差、变异过程，算法收敛速度快、结构简单。但PSO也同时存在着容易早熟收敛、搜索精度较低、后期迭代效率不高等缺点。

对于小样本问题，学者们普遍采用半监督学习和集成学习两种范式对SVM进行改进。半监督学习[12]利用未标记样本所隐含的地物类型在特征空间中的结构信息，拟合出一个更有代表性的分类器；集成学习[13]综合多个同构或异构学习机对同一个问题进行学习，进而提高分类器的泛化能力。然而两种方法的发展几乎是并行的，最近一项研究结果表明，集成学习与半监督学习之间存在许多互补性，且二者的混合范式(即半监督集成)可以更大程度地改进学习系统的泛化能力[14]。因此，设计有效的半监督集成方案是处理小样本问题的另一个崭新思路[15]。

鉴于上述分析，本文提出一种新的半监督集成SVM分类模型，该模型可以同时解决SVM参数选择不准确和小样本问题，并在多光谱遥感影像的土地覆盖分类实验中获得很好的分类效果。

2 半监督集成SVM的分类模型构建

Krogh和Vedelsby[16]以回归学习集成推导的泛化误差公式表示为：

$$E=\bar{E}-\bar{A} \tag{1}$$

式中，$\bar{E}$表示个体分类器固有误差，$\bar{A}$表示个体分类器之间差异。该公式表明，要获得好的集成，就需要降低个体分类器的误差并增加个体分类器之间的差异。因此本文从个体生成(使用程序来生成个体分类器)和结论生成(选择特定的策略来组合分类器)两个部分考虑，提出半监督集成SVM分类策略。具体技术路线如下：(1)个体生成部分一方面利用自适应变异粒子群算法(Self-adaptive Mutation PSO，SAMPSO)优化SVM分类器参数以获得高精度分类器个体，另一方面采用Gustafson-kessel(GKclust)模糊聚类算法控制Self-training错误标记样本的加入以提高个体分类器的差异性；(2)结论生成部分采用加权投票策略将半监督分类器个体集成。具体描述如图1所示。

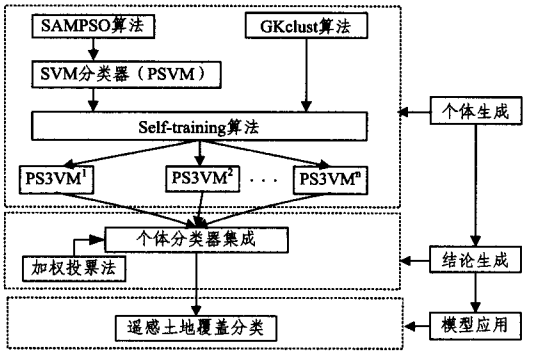

图1 半监督集成SVM分类模型技术路线

2.1 个体生成算法

2.1.1 基于自适应变异粒子群的SVM参数优化分类算法(PSVM)

SVM的主要思想是建立一个超平面作为决策曲面，使得正例和反例之间的隔离边缘(Margin width)被最大化。寻找最优超平面即正反例间隔最大化问题，最终归结为一个二次规划问题。

假设给定训练样本集$\{(x_i,y_i),i=1,2,\cdots,n\}$由两类组成，其中，$x_i\in R^n$为$N$维向量，$y_i\in\{+1,-1\}$。SVM通过解决式(2)所示的优化问题获得理想的分类超平面决策函数$f(x)=(w\cdot x)+b$，其中ω和b分别为权向量和偏移量。

$$\begin{cases}\min \dfrac{1}{2}\|\omega\|^2+c\sum\limits_{i=1}^{n}\xi_i \\ y_i(\omega\cdot x_i+b)\geqslant 1-\xi_i,\xi_i\geqslant 0,i=1,2,\cdots,n\end{cases} \tag{2}$$

式中，c为惩罚系数，控制对错分样本的惩罚程度，c越大表明对错误分类的惩罚越大。当训练样本为非线性时，采用适当核函数就可代替向高维空间的非线性映射，从而实现某一非线性变换后的线性分类。建立lagrange函数求得其对偶式，最终得到的分类决策函数为式(3)所示。

$$\begin{aligned}f(x)&=\mathrm{sgn}\{\omega\cdot\varphi(x)+b\}\\&=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i^*y_iK(x_i,x)+b^*\}\end{aligned} \tag{3}$$

在SVM 4种核函数中应用最广泛的是径向基核(Radial Basis Function，RBF)核，如式(4)所示，RBF核具有较宽的收敛性，不受维数以及样本数量的严格限制，是较为理想的分类依据函数。

$$K(x_i,x)=\exp(-\frac{1}{\sigma^2}|x_i-x|^2) \tag{4}$$

本文SVM的主要参数，除了惩罚系数c，还包括RBF核函数参数γ。粒子群算法(PSO)是常见的SVM参数寻优方法，然而，PSO算法运行过程中容易陷入局部最优，即出现所谓的早熟收敛现象，这种缺点会造成SVM分类参数寻找不准确，进而影响分类精度。Liu[17]提出利用自适应变异粒子群算法优化SVM分类参数，算法通过分析早熟收敛的判定准则确定算法是否达到全局收敛，当出现早熟收敛时，通过变异算子改变粒子的前进方向，从而让粒子进入其它区域进行搜索，在其后的搜索过程中，算法就可能发现新的个体极值以及全局极值。Liu将所提出的模型应用于遥感影像分类，并与传统PSO优化SVM参数比较，实验表明所提模型可以有效提高SVM分类精度，详细算法参考文献[17]。

2.1.2 自训练半监督分类算法(PS3VM)

自训练方法是半监督学习比较常用的方法，其基本思想是首先用少量标签样本训练一个分类器，然后用此分类器对

所有无标签样本进行分类，并给每个无标签样本标上类别标签和相应的置信度；再将置信度高的样本连同它的类别标签合并到训练集中继续训练分类器；重复上述过程，直至结束条件满足。自训练方法利用数量很少的标签样本获得初始弱分类器，然后利用弱分类器估计无标签样本的标签置信度，很有可能将错误类别加到训练集中，以致在迭代的过程中这个错误累积并加强。Maulik 等[18]曾提出利用 FCM 模糊聚类算法筛选标注对象，以控制错误样本的加入，然而 FCM 不考虑图像上下文中的任何空间信息，使得它对噪声非常敏感；此外 FCM 算法采用平方误差和准则，仅适合于发现球形或类似球形分布的类别，这些很难满足趋于超椭球体分布的遥感数据分类。鉴于此，本文在 PS3VM 模型中引入 GKclust 模糊聚类算法，GKclust 算法是距离自适应动态聚类算法的模糊推广，可以有效地搜索超椭球、平面或线型的数据类别[19]（限于篇幅，GKclust 算法从略）。

PS3VM 算法的主要思想：在未标签样本标注过程中，利用 GKclust 对未标签样本产生模糊隶属度函数，然后通过判定模糊隶属度函数值的大小确定是否将其作为标注对象，从而远离无效标签样本值。PS3VM 算法如表 1 所列。

表 1　PS3VM 算法

步骤	执行内容
Step 1	初始化标签样本集 T=L，无标签样本集 M，$\tau=\tau_0$。
Setp 2	当 M≠Φ 执行如下操作。
Step 3	利用标签集训练 SVM，并利用自适应变异 PSO 进行参数优化，构建初始分类器。
Step 4	在集合 T 中利用 Gkclust 模糊聚类算法产生聚类中心 V。
Step 5	以 V 为初始聚类中心，在无标签集合中生成无标签样本的模糊隶属度函数值。
Step 6	将隶属度高的样本点组成候选集合 N。
Step 7	利用 PSVM 对 N 进行标注。
Step 8	基于 τ 产生标签子集 ψ。
Step 9	更新标签集 T←T∪ψ。
Step 10	更新无标签集 M←M−ψ。
Step 11	如果 ψ=Φ 降低 τ 的值。
Step 12	判断循环是否结束。
Step 13	利用 T 再次训练 PSVM。

其中 L 为标签样本集，M 为无标签样本集。首先利用模糊聚类算法 Gkclust 对初始标签样本点 L 进行非监督聚类，产生 H 个类别的聚类中心 V；然后以 V 为初始聚类中心，利用 Gkclust 对无标签样本点 M 进行非监督聚类，产生 H 个类簇和所有无标签样本点的模糊隶属度函数 u_i，在各个类簇中，将距离聚类中心较近的点（模糊隶属度函数 u_i 较高的点）作为未标签样本的候选集合 N；利用 PSVM 模型，同时设定一个阈值 τ 对候选集进行样本标注形成集合 ψ；接下来将 ψ 增加到 L 中，并在 M 中将 ψ 删除，如此迭代，直至 M 为空。

2.2　结论生成算法

2.2.1　加权投票法

集成学习的主要思想是利用分类器的融合改善单个分类器的不足，其性能一方面取决于多样性强的个体学习器，另一方面依赖于成员分类器的有效组合。常用的组合方法有叠加法、选择法、投票法。对于分类问题，通常采用投票法[20]，其中包括多数投票和加权投票法。

加权投票法是将每个成员分类器均赋予一定的权重，权重通常通过在训练集上测量每个成员分类器精度获得；且权重与精度成正比，即分类能力好的基分类器被赋予较大的权系数；而分类能力相对差的基分类器赋予较小的权系数，集成的结果取决于加权和，而多数投票实际是加权投票的特例，即权重值相等的加权投票法。设 $h_t(t=1,2,\cdots,T)$ 是第 t 个成员分类器的决策函数，$w_t(t=1,2,\cdots,T)$ 是相应的权重，则最后的决策如式(5)所示。

$$f(x)=\mathrm{sign}\left(\sum_{i=1}^{T} w_i h_i(x)\right) \tag{5}$$

2.2.2　半监督集成分类算法(EPS3VM)

PS3VM 半监督方法利用未标记数据有效地应对训练样本的不足，同时也产生若干性能差异的个体分类器。接下来利用加权投票法将这些个体分类器集成，以进一步提高分类模型的泛化能力。

假设利用自训练算法产生 T 个分类器个体 $S^1,S^2,\cdots,S^T$，一幅遥感影像分类问题包含 C 个类别。

算法步骤见表 2。

表 2　EPS3VM 算法

步骤	执行内容
Step 1	将各基分类器 $S^1,S^2,\cdots,S^T$ 的分类混淆矩阵获得的各类别的用户精度作为权值 W_j^i(i=1,2,…,T;j =1,2,…,C)；
Setp 2	各基分类器对未知像元 X 分类后，将分类结果相同的各基分类器对该类别的权值相加，即得到像元 X 属于各类别的权值之和 $\sum_{i=1}^{T} W_j^i$(j=1,2,…,H)；
Step 3	比较权值之和的大小，将最大值对应的类别作为像元 X 的最终类别标签。

3　半监督集成 SVM 模型在土地覆盖分类中的应用

为了测试 EPS3VM 分类模型的性能，将半监督集成模型应用于多光谱遥感影像的土地覆盖分类实验，同时与 PSVM，PS3VM 进行对比。

3.1　实验数据

本文选择 2006 年 9 月 22 日获取的行列号为 115-30 的多光谱 Landsat-5 TM 遥感影像（30 米空间分辨率，UTM 投影）。根据植被的光谱特征和空间分布规律，本文提取了 8 个特征，包括 TM 图像的 6 个波段(1-5,7)、K-T 变换的第一主分量、植被指数(NDVI)。热红外波段 TM6 的被排除，因为它所包含的植被分类信息较少。

根据研究区实际情况并参考全国土地利用分类系统、东北植被分布图，将实验区分为 5 个土地利用类型，即林地、水体、农田、住宅、裸地。为了保证每个类别数据的变化性和代表性，数字集采用随机像素的选择策略。土地覆盖类型及样本数量如表 3 所列。

表 3　类别及样本数量

类别代号	类别名称	样本
ω_1	林地	334
ω_2	水体	229
ω_3	耕地	268
ω_4	住宅	190
ω_5	裸地	229
类别及样本总数	5	1250

3.2　结果与精度分析

当训练样本较少时，未标记样本参与的半监督分类方法可有效提高分类精度，但随着已标记分类样本的增加，未标记样本的作用越来越小。首先为了更好地体现小样本的特点，将随机抽取的少部分样本作为训练样本（占每类样本的 30%），整个数据集用于测试。分别采用 PSVM，PS3VM，EPS3VM 3 种分类模型进行对比实验，将分类精度、Kappa 系

数及相应的参数值列于表 4。实验结果表明，使用 EPS3VM 方法分类得到的分类精度比 PS3VM 模型高出 4.72%，比 PSVM 模型高出 8.4%，Kappa 系数也要高于 PS3VM 模型 0.0596，高于 PSVM 模型 0.106。表 5 显示 3 种方法不同类别的混淆矩阵，实验结果均证实 EPS3VM 能有效提高影像的分类精度。

表 4　分类参数、分类精度和 Kappa 系数的比较

分类模型	惩罚参数 c	核函数参数 γ	分类精度(%)	Kappa 系数
PSVM	52.6312	27.1031	88.48	0.8546
PS3VM	24.0602	12.8923	92.16	0.9010
EPS3VM	35.6322	28.1043	96.88	0.9606

表 5　3 种方法的混淆矩阵

PSVM	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	334	0	0	0	0
ω_2	0	169	0	0	60
ω_3	5	2	238	3	20
ω_4	0	0	15	166	9
ω_5	3	0	0	27	199
PS3VM	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	334	0	0	0	0
ω_2	0	194	0	0	35
ω_3	3	2	251	4	8
ω_4	0	0	12	170	8
ω_5	5	0	0	21	203
EPS3VM	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	334	0	0	0	0
ω_2	0	229	0	0	0
ω_3	1	0	263	1	3
ω_4	0	0	12	172	6
ω_5	3	0	0	13	213

上述实验利用遥感影像的数字集，有效地测试了 EPS3VM 的性能。在本实验中，将 3 种分类模型应用于覆盖研究区域 115-30TM 影像子集的分类并产生分类专题图，分类结果如图 2 所示。由于研究区主要土地覆盖类型为植被，因此图 2(a)为研究区 5,4,3 波段合成图；图 2(b)显示 PSVM 的分类结果，其中 1250 个样本点作为训练集；图 2(c)和图 2(d)以 1250 样本点作为已知样本点，并随机从影像上搜集 3000 个未知标签点进行半监督分类和半监督集成分类。从分类图中可以得出结论如下：首先，研究区域的主要土地覆盖类型为林地，同时可以看出森林正面临着快速城市化的威胁，为了有效地保护森林资源，有必要对此研究区进行动态监测。其次，比较 3 种分类图，不难发现 PSVM 分类模型在林地和耕地存在错分现象，PS3VM 的主要问题在于耕地、裸地和住宅分类错误，而 EPS3VM 在遥感影像数据上的分类明显优于其它两种方法。

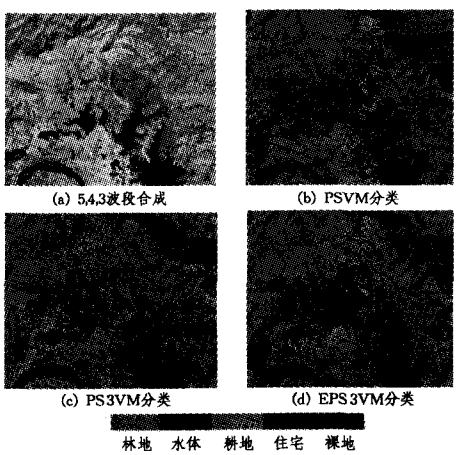

图 2　3 种分类方法对 TM 影像子集产生的分类专题图

结束语　本文针对 SVM 遥感分类中存在的参数选择和小样本问题，提出基于半监督集成 SVM 分类方法。该方法有如下特点：1)以 PSVM 模型作为基分类器，利用 Self-training算法产生半监督分类器个体，其中在半监督学习过程中为了避免错误样本的加入，引入 GKclust 模糊聚类算法；2)半监督学习和集成学习两种范式的结合，一方面充分利用大量廉价的未标记样本来减少对有标记样本的需求量；另一方面，未标记数据能够增加个体分类器之间的差异性，从而进一步提高学习系统的泛化能力；3)利用所提模型解决遥感土地覆盖分类的实验表明，在相同样本数量条件下相比于其它分类技术，该模型能获取更丰富、更准确的遥感类别信息。

同时，本文算法亦存在一些待改进之处。例如，在个体分类器集成部分，对权系数产生的问题上，可采用更有效的策略。

参 考 文 献

[1] 林剑，钟迎春，彭顺喜，等. 多光谱遥感图像土地利用分类区域多中心方法[J]. 遥感学报，2010，14(1)：173-179

[2] 韩敏，林晓峰. 一种基于 Wedgelet 变换的遥感图像分类算法[J]. 红外与毫米波学报，2008，27(4)：280-284

[3] Knorn J, Rabe A, Radeloff V C, et al. Land cover mapping of large areas using chain classification of neighboring Landsat satellite images[J]. Remote Sensing of Environment, 2009, 113 (5): 957-964

[4] Heikkinen V, Tokola T, Parkkinen J, et al. Simulated multispectral imagery for tree species classification using support vector machines[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(3): 1355-1364

[5] Lardeux C, Frison P L, Tison C, et al. Support vectormachine formultifrequency SAR polarimetric data classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2009, 47(12): 4143-4152

[6] Huang Xin, Zhang Lian-pei. Road centreline extraction from high-resolution imagery based on multiscale structural features and support vector machines[J]. International Journal of Remote Sensing, 2009, 30(8): 1977-1987

[7] 冯晓毅，王西博，王雷，等. 基于改进 JSEG 算法的高分辨率遥感图像分割方法[J]. 计算机科学，2012，39(8)：284-287

[8] 戴宏亮，戴道清. 基于 ETAFSVM 的高光谱遥感图像自动波段选择和分类[J]. 计算机科学，2009，36(4)：268-272

[9] LaValle S M, Branicky M S. On the relationship between classical grid search and probabilistic roadmaps [J]. International Journal of Robotics Research, 2002, 23(8): 673-692

[10] Fröhlich H, Chapelle O. Feature selection for support vector machines by means of genetic algorithms[C]// Proceedings of the 15th IEEE international conference on tools with artificial intelligence. USA, Sacramento, C A, 2003

[11] Huang Cheng-lung, Dun Jian-fan. A distributed PSO-SVM hybrid system with feature selection and parameter optimization [J]. Applied Soft Computing, 2008, 8: 1381-1931

[12] Zhou Zhi-hua, Li Ming. Semi-supervised learning by disagreement[J]. Knowledge and Information Systems, 2010, 24(3): 415-439

[13] Zhou Zhi-hua. Ensemble learning[M]// Li S Z, ed. Encyclopedia of Biometrics, Berlin: Springer, 2009, 270-273

[14] 邬俊，段晶，鲁明羽. 基于偏袒性半监督集成的 SVM 主动反馈方案[J]. 模式识别与人工智能，2010，23(6)：745-751

[15] Wu Jun, Lin Zheng-kui, Lu Ming-yu. Asymmetric Semi-Supervised Boosting for SVM Active Learning in CBIR[C]// Proceedings of the ACM International Conference on Image and Video Retrieval. Xi'an, China, 2010

[16] Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning[J]. Advances in Neural Information Processing Systems, 1995(7): 231-238

[17] Liu Ying, Zhang Bai, Huang Li-hua, et al. A novel optimization parameters of support vector machines model for the land use/cover classification[J]. International Journal of Food, Agriculture & Environment, 2012, 10(2): 132-138

[18] Maulik U, Chakraborty D. A self-trained ensemble with semisupervised SVM: An application to pixel classification of remote sensing imagery[J]. Pattern Recognition, 2011, 44: 615-623

[19] 黄金杰，李士勇，蔡云泽. 一种建立粗糙数据模型的监督模糊聚类方法[J]. 软件学报，2005，16(6)：744-753

[20] 单丹丹，杜培军，夏俊士. 基于多分类器集成的"北京一号"小卫星遥感影像分类研究[J]. 遥感应用，2011，2：69-78

(上接第 191 页)

趋向于 1(一般大于 0.9)，不同人的人脸数据的 $r(x)$ 都较小并趋向于 0。CCMEBTL 通过迁移学习尽量多地利用原有信息实现了领域自适应，我们可以做到对人脸的有效识别。

同时通过 CCMEBTL 算法，在 3.3 节表 1 中可求得训练集(源域空间 D1)的最小包含球球心 c 的坐标和测试集(目标域域空间 D2)最小包含球球心 C1 的坐标，继而可求出不同人脸图像的球间距离，如表 7 所列。球间距离越小，表示源域与目标域相似度越高。

表 7 不同人脸图像的球间距离

训练集/测试集	1.bmp	2.bmp	3.bmp	4.bmp	5.bmp	6.bmp	7.bmp	8.bmp
1.bmp	9.53 e−14	4.57	3.44	4.72	11.10	13.79	12.86	14.87
2.bmp	4.5705	9.74 e−14	2.24	1.26	10.92	12.42	13.20	16.52
3.bmp	3.44	2.24	9.55 e−14	2.51	10.83	12.70	12.94	15.68
4.bmp	4.72	1.26	2.51	9.71 e−14	11.59	13.11	13.94	16.93
5.bmp	11.10	10.92	10.83	11.59	9.23 e−14	5.16	5.11	7.74
6.bmp	13.79	12.42	12.70	13.11	5.16	9.00 e−14	6.38	8.70
7.bmp	12.86	13.20	12.94	13.94	4.70	6.38	8.87 e−14	8.32
8.bmp	14.87	16.52	15.68	16.93	7.74	8.70	8.32	0.10e−14

通过表 7 观察可以发现，两大类数据内部子集的球心间距明显小于不同类子集之间的球心间距，即同一人的人脸图像球间距明显小于不同人的人脸图像球间距。结果显示算法能较好地体现不同领域之间的相关性，具有较好的领域自适应性。

结束语 本文将 MEB、CCMEB 理论应用在迁移学习研究上，提出了 MEBTL 算法和 CCMEBTL 算法。在求解目标域球心位置时尽可能多地利用到源域数据完成知识传递，并发现不同域之间的内部联系。最后通过比较不同域的概率统计比可实现数据的修正和校正。为了满足大样本数据集运算要求，引入了 CVM、CCMEB 理论。大量的实验内容验证了本文算法的有效性和快速性。应当指出本文算法仍有可深入研究之处，如何将其应用于数据分类和数据回归将是我们下一步的研究重点。

参 考 文 献

[1] Dai W, Yang Q, Xue G, et al. Boostimg for transerfer learning [C]// Proceedings of the 24th International Conference on Machine Learning. USA Corvasllis: ACM, 2007: 193-200

[2] Pan S J, Yang Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359

[3] Blitzer J, McDonald R, Percira F. Domain adaptation with structural correspondence learning [C]// Processings of the 2006 Conference on Empirical Methods in Natural Language Processing. PA USA: SIAM, 2006: 120-128

[4] Hal Daum'e III. Daniel Mareu Domain adaptation for statistical classifiers[J]. Journal of Artificial Intelligence Research, 2006, 26(4): 101-126

[5] Blitzer J, Crammer K, Kulesza A, et al. Learning bounds for domain adaptation[C]// Proceedings of the 21st Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT, 2008: 129-136

[6] Dai W, Xue G, Yang Q, et al. Co-clustering based classification for out-of-domain documents[C]// Proceedings of 13th ACM SIGKDD. New York: ACM, 2007: 210-219

[7] Tax D M J, Duin R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11): 1191-1199

[8] Tsang I, Kwok J, Zurada J. Generlized core vector machines [J]. IEEE Transactions on Neural Networks, 2006, 17(5): 1126-1139

[9] Tsang I, Kwok J, Cheung P. Core vector machines: Fast SVM training on very large data sets [J]. Journal of Machine Learning Research, 2005, 6(4): 363-392

[10] Fang S-H, Lin T-N. Indoor location system based on discriminant-adaptive neural network in IEEE 802.11 environments [J]. IEEE Transactions on Neural Networks, 2008, 19(11): 1973-1978

[11] Yang Q, Pan S J, Zheng V W. Estimating location using Wi-Fi [J]. Intelligent Systems, IEEE, 2008, 23(1): 8-13

[12] Satpal S, Sarawagi S. Domain Adaptation of Conditional Probability Models via Feature Subsetting [C]// Proceedings of PKDD. Heidelberg: Springer-Verlag Press, 2007, 4702: 224-235

[13] Pan S J, Kwok J T, Yang Q. Transfer learning via dimensionality reduction[C]// Proc. 23rd Assoc. for the Advancement of Artificial Intelligence (AAAI) Conf. Artificial Intelligence. Chicago, IL, July 2008: 677-682

[14] Pan S J, Tsang I W, Kwok J T, et al. Domain Adaptation via Transfer Component Analysis [J]. IEEE Transactions on Neural Networks, 2011, 22(2): 199-210

[15] Osuna E, Freund R, Giroal F. Training support vector machines: an application to face diction[C]// Prceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Juan, 1997: 130-136