中图法分类号: TP751 文献标志码: A 文章编号: 1006-8961(2011)11-2024-06
论文索引信息:韩敏,朱新荣. 应用于地物识别的改进轮转森林算法[J]. 中国图象图形学报,2011,16(11):2024-2029

应用于地物识别的改进轮转森林算法

韩敏,朱新荣
(大连理工大学电子信息与电气工程学部, 大连 116024)

摘 要: 针对数量激增、数据类型复杂的遥感影像,准确和具有普适性的分类是亟待解决的问题。提出一种轮转径向基函数神经网络模型应用于遥感影像的处理方法。通过对输入数据的特征变换,使特征总集变为多个子特征集,依据 PCA(主成分分析)变换处理这些新的子特征集,将得到的系数用于改变训练样本,增加基分类器之间的差异度,提高分类精度。以扎龙湿地为研究对象将该算法与其他方法比较,结果显示本文方法能得到更准确的分类结果,而且具有较高的泛化精度以及较小的过学习现象。

关键词: 混合算法;径向基函数神经网络;轮转森林;特征变换

Modified rotation forest algorithm used for remote sensing image classification

Han Min, Zhu Xinrong
(*Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian* 116024 *China*)

Abstract: The amount of remotely sensed data increases rapidly, and the information contained in this data becomes more and more complicated, the way how to classify these datasets generalized and effectively is a problem which needs urgently to be solved. A modified rotation forest algorithm is proposed which takes the RBFNN as the base classifier to classify the remote sensing image. The input training dataset is changed by the rotation forest which can output a much small sub-feature. Then the non-redundancy feature set is got by using PCA technology to process these new sub-features. Finally, the training dataset changes according to the coefficient by the PCA transformation. This change will lead a higher diversity factor among these sub-classifiers which will give a much higher accuracy. The proposed method can obtain higher classification accuracy than other traditional methods when it used on the Zhalong wetland remote sensing image, and this algorithm has much higher generalization ability and much less over study phenomenon.

Keywords: hybrid algorithm; RBFNN; rotation forest; feature transformation

0 引 言

面对现今用来进行地物识别的遥感数据的数量以及种类越来越多,需要进行处理实际问题也越来越复杂的情况[1-2],单纯地对一种算法进行改进已经不能满足解决问题的需要,然而集成学习方法可以较好地解决该问题[3-6]。集成学习方法按照分类器之间的种类关系可以分为异态集成学习和同态集成学习两种。异态集成学习指的是使用各种不同的分类器进行集成,它的优势在于异态集成学习中的某种基本算法会对某类特定数据样本比其他的基本算法更为有效,得到的效果也会更好;同态集成学习是指集成的基分类器都是同一类分类器,只是基分类器之间的参数有所不同,它针对各种不同的数据类型用抽样与集成进行结合,对原始训练集进行一系列抽样,产生多

收稿日期:2010-09-14;**修回日期**:2010-12-01
基金项目:国家自然科学基金项目(61074096);国家科技支撑计划项目(2006BAB14B05)。
第一作者简介:韩敏(1959—),女,教授。1999 年于日本国立九州大学获工学博士学位,研究方向为神经网络、混沌序列分析以及其在控制和识别方面的应用。E-mail: minhan@dlut.edu.cn。

个分类器,然后用投票或合并的方式输出最终结果。

决策树的研究为这两种集成方式提供了研究平台,利用决策树来进行特征提取也是其中的一个研究方向,然而有学者提出决策树在构建过程中进行特征选择有不足之处[7]:决策树中出现的特征是有等级的,浅节点的特征明显要比深节点处的特征更为重要,如果直接综合每个节点的特征来进行特征选择,就完全忽略了这种等级特性,有可能使特征子集的潜在作用得不到发挥。因此如何正确地使用决策树来进行特征选择也是一个研究热点,这也是轮转森林(Rot-F)[8]提出的出发点。轮转森林对特征进行深层划分,将原特征集随机分为多个小的特征子集,利用PCA(主成分分析)变换得到的系数对原数据集进行改变,以此来建立树的过程要明显优于普通决策树在分叉过程中对原始数据的分类性能。然而决策树对遥感数据的分类往往会出现过分类的现象,所以考虑将具有自学习和自组织能力的径向基函数神经网络(RBFNN)与之进行结合,用轮转森林转换各子集后的数据样本作为神经网络的输入,对应转换后的样本类标作为网络的输出,以此来构造多个子分类器集。通过将原数据集进行分化的方法,构造多个内部参数差异较大的同态集成基分类器,提升总体的分类精度,集成两种方法的优势达到对遥感地物的识别更加精确。

1　Rot-F 和 RBFNN 的集成算法

轮转径向基函数神经网络首先是利用轮转森林对特征集的转换,得到多个新的特征子集,并将改变后的数据用于 RBFNN 的分类过程。对原始数据集的处理过程如下:将原始数据按照特征集随机分为多个小的特征子集数据块,之后在每个特征子集的数据块中依据数据实例进行重采样,将重采样后的数据小块进行主成分分析,使得到的各特征值的重要性程度系数与原数据集进行相应的乘积变动。因采样过程得到的数据块要小于原始数据块,所以得到特征集的大小也就不同,以新特征集作为各个 RBFNN 基分类器的输入来训练模型和预测数据。具体过程描述如下:

令 $\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$ 表示一组 n 维的数据样本点,令 $\boldsymbol{X}$ 表示一个维数为 $N\times n$ 的训练样本集;$\boldsymbol{L}$ 表示该数据集中每个实例所对应的类标,表示为 $\boldsymbol{L}=[l_1,l_2,\cdots,l_N]^{\mathrm{T}}$,其中 l_i 表示所有类标 $\{w_1,w_2,\cdots,w_c\}$ 中的一个类标号。定义 $D_1,\cdots,D_p$ 表示集成过程中的子分类器,F 表示这些子分类器建立时所依据的特征总集的集合。为了利用训练样本集来构建这些子分类器,需要进行以下几个步骤:

1)随机的将 $\boldsymbol{F}$ 分为 K 个子集,这些子集可能是相对独立的,也有可能是相互交叉重复的。为了最大可能地增加多样性,选用相对独立的特征子集。同时为简化计算,假定 K 是一个可调的变量,那么对于每个特征子集所包含的特征个数为 $M=n/K$。

2)定义 $\boldsymbol{F}_{i,j}$ 为 D_i 个子分类器用来进行训练的第 j 个特征子集。利用 Bootstrap 进行随机抽样选取一组数据样本,抽样的比例为该子集数据的 75%。仅利用 $\boldsymbol{F}_{i,j}$ 中的 M 个特征进行 PCA 变换,将得到的主成分系数 $a_{i,j}^{(1)},\cdots,a_{i,j}^{(M_j)}$ 进行存储,每个系数的维数为 $M\times 1$。对于上述的矩阵,得到的特征值有可能是 0,所以可能得不到所有的 M 个向量,即 $M_j\leqslant M$。因此通过在这些子集中运用 PCA 技术而不是在所有的数据集中应用,其目的就是为了避免在不同的子分类器中对于相似特征子集产生类似的系数。

3)将所获得的带有主成分系数的向量进行综合,构造一个具有稀疏性的轮转矩阵 $\boldsymbol{A}_i$,其表示如下:

$$\boldsymbol{A}_i=\begin{bmatrix}\boldsymbol{r}_1 & [\boldsymbol{0}] & \cdots & [\boldsymbol{0}]\\ [\boldsymbol{0}] & \boldsymbol{r}_2 & \cdots & [\boldsymbol{0}]\\ \vdots & \vdots & & \vdots\\ [\boldsymbol{0}] & [\boldsymbol{0}] & \cdots & \boldsymbol{r}_K\end{bmatrix} \tag{1}$$

矩阵中用 $\boldsymbol{r}_j=[a_{i,j}^{(1)},a_{i,j}^{(2)},\cdots,a_{i,j}^{(M_j)}]$ 表示第 j 个特征子集变换得到的主成分系数,其中:$i=1,2,\cdots,p$;$j=1,2,\cdots,K$;$[\boldsymbol{0}]$ 表示与 $\boldsymbol{r}_j$ 维数相同的零矩阵。为了利用训练集来构造子分类器 D_i,首先对 $\boldsymbol{A}_i$ 的各列进行重新排列,使得排列之后的系数与原特征集的排列相一致。定义经过重新排列后的轮转矩阵为 $\boldsymbol{A}_i^*$,矩阵维数为 $n\times N$。

4)用来进行训练 D_i 个子分类器的训练数据集为 $\boldsymbol{X}\boldsymbol{A}_i^*$,将这样的训练集作为输入对 RBFNN 进行训练来建立神经网络模型,训练部分分为两个阶段:

(1)训练输入层和隐含层之间的径向基函数,即确定基函数的中心和方差。隐含层的径向基函数有多种形式,最常用的是高斯函数。设初始聚类中心为 $\boldsymbol{t}_g$,初始聚类中心的个数为 G,则高斯函数为

$$\phi(\boldsymbol{X}\boldsymbol{A}_i^*,\boldsymbol{t}_g)=\exp\left(-\frac{\|\boldsymbol{X}\boldsymbol{A}_i^*-\boldsymbol{t}_g\|^2}{2\sigma_g^2}\right) \tag{2}$$

$$g=1,2,\cdots,G$$

训练的目的是确定隐含层节点的个数及中心向

量 $\boldsymbol{c}_g$ 和宽度 σ_g。采用 K-均值聚类方法确定中心向量 $\boldsymbol{c}_g$，可得宽度 σ_g 为

$$\sigma_1 = \sigma_2 = \cdots = \sigma_G = \frac{d_{\max}}{\sqrt{2G}} \tag{3}$$

式中 $d_{\max}$ 为 $\boldsymbol{c}_g$ 之间的最大距离。

(2)训练隐含层和输出层之间的线性权值。当网络输入训练样本 $\boldsymbol{X}$ 时，网络第 j 个输出单元的实际输出为

$$y_j = w_{0j} + \sum_{g=1}^{G} w_{gj}\phi(\boldsymbol{X}\boldsymbol{A}_i^*, \boldsymbol{t}_g) \tag{4}$$
$$j = 1,2,\cdots,c$$

网络输出节点个数等于分类类别数 c，权值 $w_{gj}(g = 0,1,2,\cdots,G; j = 1,2,\cdots,c)$ 的学习通过最小均方误差(MSE)代价函数采用迭代法则进行调整并最终达到稳定。通过采样和样本聚类生成 RBFNN 的中间节点参数，中间层状态确定法选用 K-means 聚类，该方法采用最小平方误差和作为分类准则，通过迭代获得聚类中心点集，即中间层的状态，节点个数等于聚类数。

5) 该集成模型解决的是一个 c 类的模式分类问题，集成的规模为 P(即有 P 个子分类器)，各模式类分别记作，$g_1,\cdots,g_i,\cdots,g_c$，类别 i 的输出编码为[0,…,0,1,0,…,0]，即除第 i 个元素为 1 外，其余元素全为 0。通过期望的输入输出编码映射关系对 P 个成员网络进行训练。训练之后对于测试模式 Q，每一个成员网络都会给出一个输出向量，第 k 个成员网络的输出为列向量$[O_{1,k},O_{2,k},\cdots,O_{n,k}]^{\mathrm{T}}$，$k = 1,\cdots,P$，根据投票法得到该 P 个网络的集成输出如下

$$Q_{\text{ensem\_out}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_c \end{bmatrix} = \begin{bmatrix} O_{1,1} + O_{1,2} + \cdots + O_{1,P} \\ O_{2,1} + O_{2,2} + \cdots + O_{2,P} \\ \vdots \\ O_{c,1} + O_{c,2} + \cdots + O_{c,P} \end{bmatrix} \tag{5}$$

上述矩阵中各向量的组合方式，一般可分为“多数通过”，“一致通过”和“无反对票”3 种形式。因多数投票方法常被用于产生唯一类别标签的分类器，所以可认为它是将和式规则应用于后验概率 $p(\boldsymbol{c}_k \mid \boldsymbol{x}_i\boldsymbol{A}_i^*)$ 经过“硬化”后的分类器输出上的。硬化时用二值函数 Δ_{ki} 来替代 $p(\boldsymbol{c}_k \mid \boldsymbol{x}_i\boldsymbol{A}_i^*)$，如下式：

$$\Delta_{ki} = \begin{cases} 1 & p(\boldsymbol{c}_k \mid x_i\boldsymbol{A}_i^*) = \\ & \max_j p(\boldsymbol{c}_j \mid x_i\boldsymbol{A}_i^*) \\ 0 & \text{其他} \end{cases} \tag{6}$$

把样本归入到成员分类器所作预测最多的类 w_k，如下式：

$$h(x) = \arg\max_{\tilde{w}} \sum_{i=1}^{N} \Delta_{ki} \tag{7}$$

若 $\max \sum_{i=1}^{N} \Delta_{ki}/P > 0.5$ 则为“绝对多数原则”，实例被归到票数过半的类中。其中，P 为子分类器的个数，比值越高集成策略越保守。采用绝大多数投票原则，即如果 $O_{i,k} = \max_j(O_{j,k})$，则 $O_{i,k} = 1$ 且 $O_{j,k} = 0, j \neq i$。那么当 $y_j = \max_i(y_i)$ 时，模式 Q 就被划分到第 j 类(c_j)。这样通过正建立的模型，对新的输入数据进行预测得到的结果就是 Rot-RBFNN 得到的最终预测结果。

2 数据的选取及实验结果分析

实验区采用扎龙湿地，它位于黑龙江省西部，乌裕河下游齐齐哈尔市及富裕、林甸、杜蒙、泰来县交界地域，总面积为 2.1×10^9 m^2，地理坐标为 46°52′～47°32′N，123°47′～124°37′E。扎龙湿地保护区是以芦苇沼泽为主的内陆湿地和水域生态系统。它是中国北部最完整、最原始和广阔的湿地生态系统，已被列入国际重要湿地名录。该地区地物类型主要包括盐碱地、草地(包括芦苇)、水体、沼泽(轻度和重度沼泽归为一类)这些具有比较代表性的类物。选用的遥感图像为 2001-10-05 获取的 Landsat ETM + 图像，图像大小为 500×500像素，待分类图像采用 TM 影像 7 个波段中的 3、2、1 共 3 个波段的真彩色合成图像，如图 1 所示。

图 1 研究区域影像

Fig. 1 Image of study area

经目视判读,结合土地利用现状图,确定湿地影像区域内主要包括六大类地物:盐碱地、草地、水域、沼泽地、受火区域以及农用地(包括农用耕地和农用居住地)。

Landsat TM/ETM + 影像的 3、2、1 波段组合能较好地反映土地植被特征。根据目视判读结果,并且结合土地利用现状图和 2003 年 5 月、11 月的实地调查资料,在真彩色合成影像上共选取 1 700 个样本点,其中 1 020 个像素点作为轮转神经网络的训练样本,340 个像素点作为校验样本,剩余的 340 个用来测试分类精度。为进一步减小输入数据的不同所造成不同算法之间的差异性,将所有的原始输入数据都归一化到[-1,1]之间,具体结果如下表 1 所示。分别应用朴素贝叶斯(NB)方法、传统径向基函数神经网络方法(RBFNN)、普通轮转森林方法(Rotation Forest)方法[9]、文献[10]中的方法(Rot-Boost)、随机森林方法(Random Forest)[11]以及本文方法(Rot-RBFNN)对所选的区域 ETM + 遥感影像进行分类,分类结果如图 2。对应每个类标的分类结果量化后如表 2 所示。由表 2 和图 2 的 6 幅分类效果图可以看出随机森林算法对水体的分类精度要高,但是对于训练样本较少的盐碱地和草地以及农用地等分类,精度要低于本文方法,本文方法的精度分别达到了 0.917,0.6,0.795。而且从图中可以看出采用决策树作为基分类器的分类方法中出现了过分类现象,所以有较多的斑点出现,文献[10]中的方法虽然在农用地的精度上要略高一些,但是对于其他几类地物出现了明显分辨不出的情况。此外,从图 2 可以看出,本文方法虽没有 NB 方法看起来清晰、明亮,但从精度上却明显优于 NB 算法。

将上述结果采用 ROC 曲线下面积(AUC)的大小以及用误差矩阵的主要参考指标总体精度和 Kappa 系数来进行评估。其中 Kappa 系数是综合整

表 1　波段值和总体样本数目

Tab. 1　Value of bands and number of all chosen datasets

类别	TM3	TM2	TM1	样本个数	类别标号
盐碱地	60	60	60	100	1
草地	140	140	140	300	2
水体	200	200	200	300	3
沼泽	88	88	88	200	4
受火区	170	170	170	300	5
农用地	187	187	187	500	6

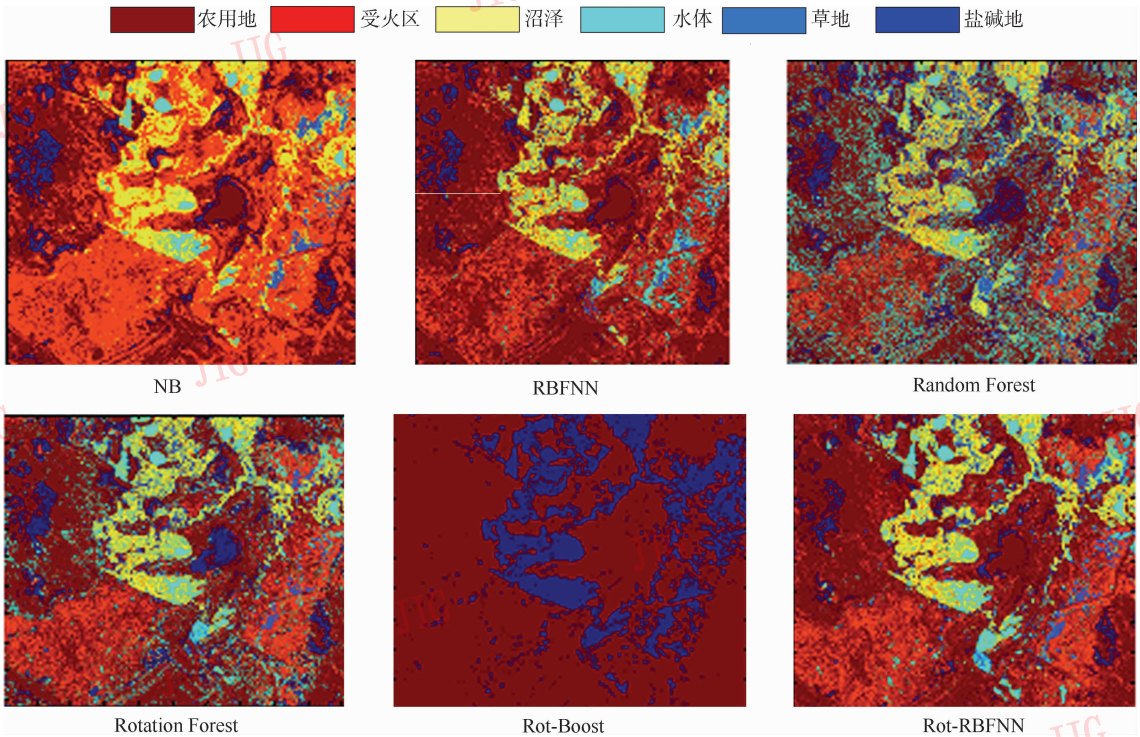

图 2　不同分类器的分类结果

Fig. 2　Classification results produced by different classifiers

表 2 不同校验样本的生产者精度对比表

Tab. 2 The comparison of producer's accuracy in different verified sample

/%

类别	NB	RBFNN	Rotation Forest[9]	Rot-Boost[10]	Random Forest[11]	本文方法
盐碱地	0.833	0.833	0.75	0	0.833	**0.917**
草地	0.796	0.796	0.852	0.852	**0.926**	0.833
水体	0.317	0.55	0.583	0	0.433	**0.6**
沼泽	0.59	0.744	0.718	0	0.667	**0.795**
受火区	**0.417**	0.117	0.3	0	0.383	0.367
农用地	0.483	0.717	0.667	**0.983**	0.717	0.833

个误差矩阵的信息提出的一个精度表达系数,可以比较不同分类器的误差矩阵在精度上的差异,得到的结果如表 4 所示。从表 3 结果可以看出,本文方法在 AUC、Kappa 系数和总体精度上都比其他几种算法高,分别达到了 0.909,0.613 8 和 68.421 1,相比于其他 5 种常用的算法,本文方法效果更好。AUC 值越大说明该分类器的分类结果越靠近于 ROC 坐标的纵轴,对应分类器的分类效果和泛化性能也就越好;Kappa 系数越高则说明对应该算法的过度拟合情况越小,也就从侧面说明该分类器所建模型的泛化能力越强。而总体精度则从准确率方面说明所提方法的有效性更强。从上述 3 个方面来说,本文方法更能给出较高的分类精度和较好的泛化性能。从算法耗时来讲,本文方法耗时稍微较长一些,达到 24.21 s。而其他几种算法的耗时基本在 10 s 左右。

为进一步验证所提算法的有效性,将该方法与其他 3 种集成算法进行横向比较,总迭代次数都为 100 次,选用的基分类器都是经过剪枝操作的 C4.5 算法,应用于 10 组 UCI 数据[12]进行仿真实验,结果见表 4。括号内的数值表示的是运行 100 次之后的平均标准偏差。从表 4 可以看出,Rot-Boost 算法

表 3 不同算法下评价指标结果比较

Tab. 3 The comparison of evaluation index among different algorithms

	NB	RBFNN	Rotation Forest[9]	Rot-Boost[10]	Random Forest[11]	本文方法
AUC	0.819	0.864	0.881	0.688	0.834	0.909
Kappa	0.419 3	0.485 5	0.532 5	0.210 5	0.541 6	0.613 8
总体精度/%	52.280 7	57.894 7	61.754 4	36.842 1	62.456 1	68.421 1

表 4 本文方法与其他集成算法在 UCI 数据中的精度比较

Tab. 4 The accuracy comparison on UCI data between the proposed method and other ensemble algorithm

/%

	Bagging	Rotboost[10]	Random Forest[11]	Rot-RBFNN
Credit_G	73.96	73.87	75.12	**75.49**
Hypothyroid	**99.52**	93.96	99.33	99.49
Diabetes	75.86	**76.81**	75.72	76.35
Heart_S	81.20	**84.06**	82.19	83.37
Hepatitis	82.20	**84.91**	83.61	84.01
Soybean	92.18	30.44	93.04	**93.69**
Zoo	42.19	60.39	93.37	**94.61**
Vowel	79.36	38.23	96.21	**96.64**
Vote	95.57	95.75	**96.34**	96.28
Vehicle	72.65	53.49	75.17	**78.30**
平均	79.47	69.19	87.01	**87.82**

虽然在 4 组数据中分类精度较高,但是同样地在其他一些数据的处理中分类精度明显降低,例如 Soybean, Vowel 等数据出现了精度在 60% 以下的情况,说明了该算法的泛化性能不强。而本文方法 Rot-RBFNN 要比前者和其余几种集成算法的泛化性能更好一些,在 10 组数据中有 5 组数据达到了最高值,而且在较多的数据中可以达到与 Random Forest 算法相似的精度,两者相差不超过 3% 。此外,从这 10 组数据的平均值来看,本文方法达到了 87.821 0% ,Random Forest 算法达到了 87.010 3% ,Bagging 算法达到了 79.469 1% ,这些结果也从侧面反映了本文算法的泛化性能较好。

3　结　论

Rot-RBFNN 集成算法具有泛化性能好、分类精度高等优势,用于地物识别的分类上可获得相对较好的分类结果。应用所提的方法,可避免决策树在遥感影像中出现的过学习现象,同时利用集成学习的方法,将神经网络与决策树进行结合,提高了模型的泛化能力以及分类精度。但是本文所建立的模型中基分类器采用的还是最基本的 RBFNN,如果采用分类精度更高的神经网络如 ESN 网络或者是在特征选择上选用 ICA 算法,对最终的结果会有进一步的提升。从目前的仿真结果可以看出,所提方法 Rot-RBFNN 与朴素贝叶斯和传统的决策树方法相比模型的泛化性能以及分类精度都有所提高。

参考文献(References)

[1] Li Shijin, Tao Jian, Wan Dingsheng, et al. Content-based remote sensing image retrieval using co-training of multiple classifiers [J]. Journal of Remote Sensing, 2010, 14(3): 500-506. [李士进,陶剑,万定生,等. 多分类器实例协同训练遥感图像检索[J]. 遥感学报,2010,14(3):500-506.]

[2] Ye Bo, Wen Yumei, He Weihua. Gait recognition based on the fusion of multiple classifiers [J]. Journal of Image and Graphics, 2009, 14(8): 1627-1637. [叶波,文玉梅,何卫华. 多分类器信息融合的步态识别算法[J]. 中国图象图形学报,2009,14(8):1627-1637.]

[3] Nicolas G P, Domingo O B. Boosting random subspace method [J]. Neural Networks, 2008, 21(9): 1344-1362.

[4] Ioannis P, Grigorios T, Ioannis V. Pruning an ensemble of classifiers via reinforcement learning [J]. Neurocomputing, 2009, 72(7-9): 1900-1909.

[5] Akhand M A H, Islam M M, Murase K. Progressive interactive training: a sequential neural network ensemble learning method [J]. Neurocomputing, 2009, 73(1-3): 260-273.

[6] Koen W D B, Kristof C, Dirk V P. Ensemble classification based on generalized additive models [J]. Computational Statistics & Data Analysis. 2010, 54(6): 1535-1546.

[7] Wang Yuanyuan, Li Jing. Analysis of feature selection and its impact on hyper-spectral data classification based on decision tree algorithm [J]. Journal of Remote Sensing, 2007, 11(1): 69-76. [王圆圆,李京. 基于决策树的高光谱数据特征选择及其对分类结果的影响分析[J]. 遥感学报,2007,11(1):69-76.]

[8] Juan J R, Kuncheva L I, Carlos J A. Rotation forest: a new classifier ensemble method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630.

[9] Liu Kunhong, Huang Deshuang. Cancer classi fication using rotation forest [J]. Computers in Biology and Medicine, 2008, 38(5): 601-610.

[10] Zhang Chunxia, Zhang Jiangshe. RotBoost: a technique for combining rotation forest and AdaBoost [J]. Pattern Recognition Letters, 2008, 29(10): 1524-1536.

[11] Leo B. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.

[12] Blake C L, Merz C J. 1998. UCI Repository of Machine Learning Databases. [DB/OL] (2010-03-01) [2010-03-12] http://www.ics.uci.Edu/～mlearn/MLR—epository.html.