文章编号: 1007-4619(2008)05-0683-09

基于误差分析的组合分类器研究

陈学泓, 陈 晋, 杨 伟, 朱 锴

(地表过程与资源生态国家重点实验室(筹), 北京师范大学 资源学院, 北京 100875)

摘 要: 提出了一种基于误差分析的组合分类器, 通过结合两种监督分类方法, 提出的算法分别估计了两种监督分类方法在计算过程中的误差, 给出了规则输出的置信区间, 再根据置信区间的大小对两种分类方法的输出结果进行加权平均, 从而得到更精确的规则输出。利用该方法对遥感图像进行分类实验, 在不同训练样本分布与不同训练样本数量的情况下, 比较新的组合分类器与单一分类器的精度。结果表明新的组合分类器能够取得比单一的分类器更高的分类精度。结果还显示出, 两个分类器的独立性越强, 组合分类器的效果越好。另外一个实验比较了新的组合分类器与和式规则组合分类器的分类精度, 结果仍显示出了新方法的优越性。

关键词: 组合分类器; 误差分析; 置信区间; 土地利用; 遥感

中图分类号: TP751.1 **文献标识码:** A

1 引 言

监督分类是遥感图像分析过程中常用的处理方法。尽管在机器学习领域不断有新的监督分类方法被提出, 但没有一种分类方法占有绝对优势。这是因为不同的分类方法有各自的缺点和优点, 某一种分类方法对特定的数据比较有效, 而对另外一组数据, 另一种分类方法可能更占优势, 这种情况被称为选择优越性(selective superiority)[1]。因此, 将不同分类器的分类结果按一定规则组合, 可以提高分类精度。组合分类器在遥感图像上的应用越来越受到关注[2], 但是基于误差分析的组合分类器还很少见, Joon Hur 和 Jong Woo Kim[3] 曾提出一种基于误差分析的组合分类方法, 然而该方法仅限于二类分类情况。本文尝试提出一种新的基于误差分析的分类器组合方法, 通过组合最大似然法[4](maximum likelihood classification, MLC)支撑向量机[5](support vector machine, SVM)以获得更高的分类精度, 并将其应用于遥感图像。最大似然法是遥感图像处理中最常用的分类方法, 它基于扎实的统计理论, 概率意义清晰, 但是它的缺点也十分明显, 当样本的分布与正态偏差较大时, 分类精度会受到严重影响。支撑向量机是近年发展起来的一种分类方法, 它基于小样本统计理论, 其对小样本的稳健性已受到广泛认可[6], 然而基本的 SVM 分类器只是二类分类器, 将其推广为多类分类器还未得到很好的解决[7], 这成为影响 SVM 分类精度的重要原因。本文提出的算法估计了 MLC 和 SVM 分类方法在计算过程中的误差, 给出了规则输出的置信区间, 再根据置信区间的大小对两种分类方法的规则输出进行加权平均, 从而获得更精确的规则输出。

2 最大似然分类法和支撑向量机分类原理

2.1 最大似然分类法

最大似然分类法是典型的基于经典统计理论的监督分类方法。假设像元的光谱信息由向量 $\boldsymbol{x}=[x_1, x_2, \cdots, x_p]$ 表示, 研究区可分成 L 类。最大似然法假设同一类的地物光谱符合正态分布。首先统计每一类训练区样本在特征空间中的分布特征, 假设类别 k 的光谱均值 μ_k 和协方差 Σ_k, 通过这两个参数可以确定正态概率密度函数, 即类别 k 中出现

收稿日期: 2007-05-21; 修订日期: 2007-10-09

基金项目: 国家 863计划(编号: 2006AA12Z103)资助。

作者简介: 陈学泓(1985—), 男。硕士研究生。现从事资源环境遥感应用研究。通讯作者: 陈 晋, E-mail: chenjin@ires.cn。

x的条件概率密度 $P(x|k)$:

$$P(x|k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-(x-\mu_k)'S_k^{-1}(x-\mu_k)/2} \tag{1}$$

式中 P 表示图像的波段数。由于遥感图像获取的像元光谱信息为离散值, 可以将概率密度直接当作概率。若第 k类的先验概率为 $P(k)$, 则相应 x 归属为第 k类的概率为:

$$P(k|x) = P(k) \times P(x|k) / \sum_{k=1}^{L} [P(k) \times P(x|k)] \tag{2}$$

为后续计算方便, 对 (2)式两边取对数可得到MLC的规则输出:

$$g_k(x) = \ln P(k|x) = \ln P(k) - \frac{1}{2}\ln|S_k| - \frac{1}{2}(x-\mu_k)'S_k^{-1}(x-\mu_k) - \frac{p}{2}\ln\pi - \ln\sum_{k=1}^{L}[P(k) \times P(x|k)] \tag{3}$$

取 $i = \arg\max_k [g_k(x)]$作为像元 x 的归属类别。

2.2 支撑向量机

基本的 SVM 分类器只能用来解决两类的分类问题。其原理如下, 给定一个线性可分的训练样本

$$S = ((x_1, y_1), \cdots, (x_n, y_n))$$

式中, x_i 表示像元的光谱信息, $y_i \in \{1, -1\}$表示 x_i 的归属类别, 求解二次优化问题

$$\begin{aligned} &\text{minimize}_{w,b} \langle w \cdot w \rangle \\ &\text{subject to } y_i(\langle w \cdot x_i \rangle + b) \geqslant 1 \\ &i = 1, \cdots, n \end{aligned} \tag{4}$$

式中, w 表示权重向量, 决定超平面的方向, b 表示超平面的偏置。(4)式的意义是求得超平面 (w, b), 该超平面实现了样本的几何间隔 $1/\|w\|$最大。对于测试样本, 有输出:

$$y = \langle w \cdot x_i \rangle + b \tag{5}$$

根据 y 的正负判定测试样本属于哪一类。对于训练样本线性不可分的情况可以通过核函数扩展或间隔泛化实现。Platt[8] 提出 SVM 二类分类器的概率输出方法:

$$r_{ij} = P(y = i | y = i \text{ or } j) = \frac{1}{1 + e^{Af\hat{} + B}} \tag{6}$$

式中, r_{ij}表示在 i, j两类中, 像元属于第 i类的条件概率, $f\hat{}$为 (5)式, A、B可以通过训练样本得到。解决 SVM 多类分类问题的研究很多, 本文采用ENVI4.3引用的方法, 即 Wu等提出的根据 SVM 二类分类器概率输出估计多类概率的方法[9]。假设研究区可分成 L类, 对任意 i, j两类生成 SVM 分类器, 共可生成 $L(L-1)/2$个二类分类器。每一个分类器可以估计某像元属于第 i类的条件概率 $r_{ij} = P(y = i | y = i \text{ or } j)$。因此 r_{ij}满足:

$$r_{ij} + r_{ji} = 1, \quad r_{ij} = \frac{P_i}{P_i + P_j} \tag{7}$$

变形得到

$$r_{ji}P_i - r_{ij}P_j = 0 \tag{8}$$

式中, P_i 表示像元属于第 i类的概率。由于共有 $L(L-1)/2$个二类分类器, 所以可构建 $L(L-1)/2$形如式 (8)的方程, 待解的未知数个数为 L。在 $L > 3$的情况下, 方程的个数超过未知数的个数, 为超定方程组, 当 $L = 4$时, 方程组的形式如下:

$$\begin{vmatrix} r_{21} & -r_{12} & 0 & 0 \\ r_{31} & 0 & -r_{13} & 0 \\ r_{41} & 0 & 0 & -r_{14} \\ 0 & r_{32} & -r_{23} & 0 \\ 0 & r_{42} & 0 & -r_{24} \\ 0 & 0 & r_{43} & -r_{34} \end{vmatrix} \begin{vmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{vmatrix} = 0 \tag{9}$$

简记为:

$$ZP = 0 \tag{10}$$

该超定方程组的最优解等价于求解如下最优化问题:

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{k} \sum_{j:j\neq i} (r_{ji}P_i - r_{ij}P_j)^2 \\ &\text{subject to} \sum_{i=1}^{L} P_i = 1, P_i \geqslant 0 \end{aligned} \tag{11}$$

求解出 P_i, 即 SVM 分类方法的概率输出, 取 $k = \arg\max_i (P_i)$作为待分像元的归属类。

3 基于误差分析的组合分类方法

本文根据文献 [10], 用标准不确定度, 也就是0.317置信水平下的置信区间来表示分类器规则输出的误差大小。

3.1 最大似然法的误差分析

利用样本的均值和协方差确定条件概率密度分布 $P(x|k)$是 MLC 方法产生误差的重要原因, 样本分布对正态的偏离会影响最大似然估计结果的准确度。最大似然估计的实质是, 用样本均值代替总体均值, 用样本协方差代替总体协方差。然而实

际上, 样本均值只是总体均值的最似然值, 样本的分布情况决定了用样本均值代替总体均值的可靠程度。利用统计学, 可以更精确地估计出总体均值的置信域[11]。p 维正态分布总体均值 μ 的 $100(1-\alpha)\%$ 置信域是一椭球, 由满足下式的 μ 的集合构成:

$$n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu)\leqslant\frac{p(n-1)}{n-p}F_{p,n-p}(\alpha) \tag{12}$$

式中, $\bar{x}$ 表示样本的均值, $S=\frac{n}{n-1}\sum$, $\sum$ 表示样本的协方差矩阵, n 表示样本的数量。求得每一类均值 $\mu_i\in\{\mu_1,\mu_2,\cdots,\mu_L\}$ 的置信域后, 如果忽略协方差的误差影响, 可以进一步得到 $g(k|x)=\ln P(k|x)$ 的置信区间。当 $\Delta\mu_i=(\bar{x}_i-\mu_i)$ 不是很大时, 有如下近似:

$$\Delta g_k(x)\approx\sum_{i=1}^{L}\frac{\partial g_k(x)}{\partial\mu_i}\Delta\mu_i \tag{13}$$

式中, $g_k(x)=g(k|x)$ 表示 x 属于类别 k 的规则输出, 令

$$\Delta u_i=\frac{\partial g_k(x)}{\partial\mu_i}\Delta\mu_i,$$

Δu_i 表示了 μ_i 的扰动对 $g_k(x)$ 产生的影响, 有

$$\Delta\mu_i=\left|\frac{\partial g_k(x)}{\partial\mu_i}\right|^{-1}\Delta u_i$$

$$=\begin{cases}((1-P(i|x))\sum_i^{-1}(x-\mu_i))^{-1}\Delta u_i, & \text{若 } i=k\\ ((-P(i|x))\sum_i^{-1}(x-\mu_i))^{-1}\Delta u_i, & \text{若 } i\neq k\end{cases} \tag{14}$$

由于式(12)表示了 μ_i 的置信域, 将式(14)代入式(12)可解得 Δu_i 满足:

$$\Delta u_i^2\leqslant\begin{cases}\left|(\sum_i^{-1}(x-\mu_i))'S_i(\sum_i^{-1}(x-\mu_i))\right|\times\\ (1-P(i|x))^2\frac{p(n_i-1)}{n_i(n_i-p)}F_{p,n_i-p}(\alpha), & \text{若 } i=k\\ \left|(\sum_i^{-1}(x-\mu_i))'S_i(\sum_i^{-1}(x-\mu_i))\right|\times\\ (P(i|x))^2\frac{p(n_i-1)}{n_i(n_i-p)}F_{p,n_i-p}(\alpha), & \text{若 } i\neq k\end{cases} \tag{15}$$

求得每一个 Δu_i 后, 根据误差传递公式[10], $g_k(x)$ 的置信区间 $\Delta g_k(x)$ 满足:

$$\Delta g_k(x)^2=\sum_{i=1}^{L}\Delta u_i^2\leqslant\frac{p(n_k-1)}{n_k(n_k-p)}F_{p,n_k-p}(\alpha)(1-P(k|x))^2\left|(\sum_k^{-1}(x-\mu_k))'S_k(\sum_k^{-1}(x-\mu_k))\right|+$$
$$\sum_{i\neq k}\frac{p(n_i-1)}{n_i(n_i-p)}F_{p,n_i-p}(\alpha)P(i|x)^2\times\left|(\sum_i^{-1}(x-\mu_i))'S_i(\sum_i^{-1}(x-\mu_i))\right| \tag{16}$$

$\Delta g_k(x)$ 表示了误差大小, 其值越大意味着估计的 $g_k(x)$ 的可信度越低。最大似然法输出的规则可以写成

$$g_{km}(x)\pm\Delta g_{km}(x)。$$

其中式(13)—式(16)为作者自行推导。

3.2　支撑向量机的误差分析

SVM 二类分类器一直被认为是十分先进的分类方法, 其对小样本的稳健性得到广泛认可。然而将 SVM 推广为多类分类器一直得不到很好解决[12-14], 也是 SVM 用于分类最有可能产生误差的地方。2.2节介绍的解决 SVM 分多类的问题的方法是通过求解最优化问题(11)实现的, 也就是求解超定线性方程组 $ZP=0$(式 10)。

若解出的最优解 P 值用 $\hat{P}$ 表示, 假设 $Z\hat{P}=e$, 其中 e 为误差项满足正态分布, 则 P 的 $100(1-\alpha)\%$ 置信椭球为[11]:

$$\Delta P'Z'Z\Delta P\leqslant(r+1)s^2F_{r+1,n-r-1}(\alpha) \tag{17}$$

其中

$$\Delta P=P-\hat{P},\ s^2=e'e/(n-r-1),$$
$$n=L(L-1)/2,\ r=L$$

所以 x 属于第 k 类的概率的 $100(1-\alpha)\%$ 置信区间 ΔP_k 为 P 的置信椭球的对角元, 满足

$$\Delta P_k^2\leqslant\frac{(r+1)s^2F_{r+1,n-r-1}(\alpha)}{(Z'Z)_{kk}} \tag{18}$$

为了与 MLC 输出的判别函数具有可比性, 同样对 SVM 输出的概率求对数, 有:

$$g_{ks}(x)=\ln P_k(x) \tag{19}$$

$$\Delta g_{ks}(x)^2=\left|\frac{\partial\Delta g_{ks}(x)}{\partial P_k(x)}\Delta P_k(x)\right|^2$$
$$=\left|\frac{\Delta P_k(x)}{P_k(x)}\right|^2 \tag{20}$$

SVM 估计的规则输出结果可以写为

$$g_{ks}(x)\pm\Delta g_{ks}(x)。$$

其中式(18)—式(20)为作者自行推导。

3.3 SVM 和 MLC规则输出的组合方法

分类器规则输出的置信区间越大, 意味着该规则输出的误差越大, 可信度也越低。利用误差大小作为权重对物理量作加权平均是数据同化中常采取的办法[15]。本文根据误差大小作为权重对规则输出作加权平均, 以获得精度更高的规则输出。新的规则输出为:

$$g_k(x)=\frac{g_{ks}(x)\,\Delta g_{ks}(x)^{-2}+g_{lm}(x)\,\Delta g_{lm}(x)^{-2}}{\Delta g_{lm}(x)^{-2}+\Delta g_{ks}(x)^{-2}} \tag{21}$$

利用新的规则输出得到新的分类结果, 取 $i=\arg\max_k[g_k(x)]$作为像元 x 的归属类别。

4 实例分析

4.1 试验区概况和数据获取

试验区位于北京市颐和园周边地区, 该地区地物类型包括了常见的城市用地类型, 包括水体、绿地、建筑、裸地和土壤, 具有比较好的代表性。选用的遥感图像为 2000-04-30 获取的 Landsat ETM 图像, 选取的图像大小 169×167像素, 如图 1。

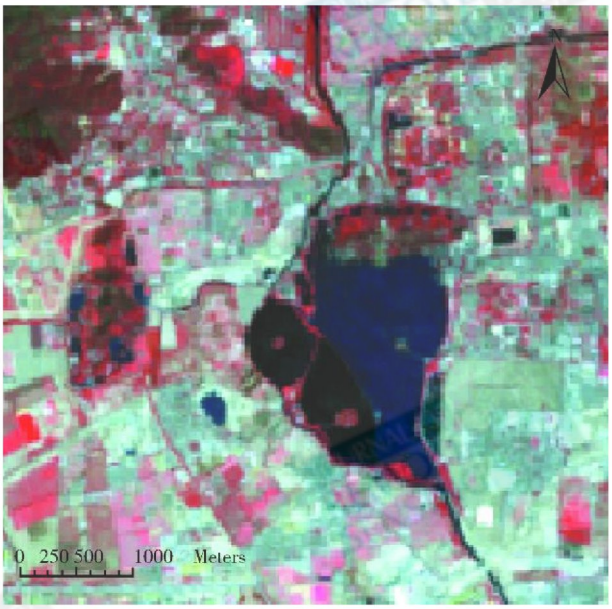

图 1 研究区域影像

Fig. 1 Image of study area

4.2 结果检验与比较

检验方法如下: (1)通过目视解译, 将该地区分为水体、绿地、建筑用地、裸地、土壤等 5类土地利用。通过野外抽样和分类后修改, 获得该地区较为准确的土地利用图, 并将其作为真实分类图 (图 2(a)); (2)根据真实分类图的结果从图像上随机选取各类别训练样本, 分别利用 SVM、MLC以及新的组合分类器进行分类; (3)比较 SVM、MLC以及新方法的分类精度。本研究中 SVM 和 MLC的各项参数设置按照 ENVI 4.3的默认设置。SVM 核函数采用径向基函数 (Gamma值取 0.617), 两者不设概率阈值。

图 2(b)、(c)、(d)分别为 MLC、SVM 和新方法的分类结果, 表 1—表 3分别为 3种分类方法的混淆矩阵。从整体的精度和 kappa系数上看, 新方法有一定程度的提升。从每一类的用户精度和生产者精度看, MLC对水体、绿地和建筑用地的错分情况比较理想, 但漏分比较多, 而对裸地和土壤而言, SVM 的错分情况比较理想, 漏分比较多。新方法的各类的用户精度和生产者精度基本能达到或接近 MLC和 SVM 中较高的精度。进一步观察可以发现, 新方法对分类精度提升较大的是建筑用地和土壤的生产者精度。这是因为, 尽管 MLC和 SVM 对两者的漏分误差差不多, 但是漏分的结构不一样 (MLC把大量的建筑用地分成裸地, 而 SVM 更多地把建筑用地分成绿地, 分成裸地的像元较少; 同样的, MLC将大量的土壤分成裸地, 而 SVM 则较多地将其分成绿地和建筑用地), 这使得错分样本集的独立性比较好, 组合分类器能够发挥出优势。新方法表现较差的类别, 如水体的生产者精度, 从混淆矩阵中可以看到两个分类器对水体的漏分结构比较相似, 错分样本集的独立性较差, 所以新方法的提升潜力有限。

4.3 不同训练样本的影响

研究表明, 相同的分类器对不同的样本有着不同的响应[16], 训练样本对分类精度的影响比分类技术对精度的影响还要大[17, 18]。因此, 考察不同训练样本下新分类器的性能, 是十分必要的。

4.3.1 不同训练样本分布的影响

训练样本位置选择的不同往往会对分类结果造成很大的影响, 本研究对该区域按 1% (样本数 282)分层随机采样 30次, 以验证 MLC, SVM 和新分类方法对样本分布差异的稳健性。图 3表示了 30次随机采样的 3种分类方法的 kappa系数对比。MLC, SVM 和新方法的 kappa系数的平均值分别为 0.732, 0.754 和 0.763。可以看到新方法在精度上有一定程度的提升。但是有几次实验, 新方法分类精度提升特别微小甚至比起 SVM 的分类精度有下降, 这是因为这几

次实验 SVM 取得了较高的分类精度, 使得组合能够提升的潜力有限。仔细观察可以看到, 如果 SVM 和 MLC两个分类器的精度比较接近时, 新方法有较明显的提升。本文定义两个参量来着重表示这个现象。

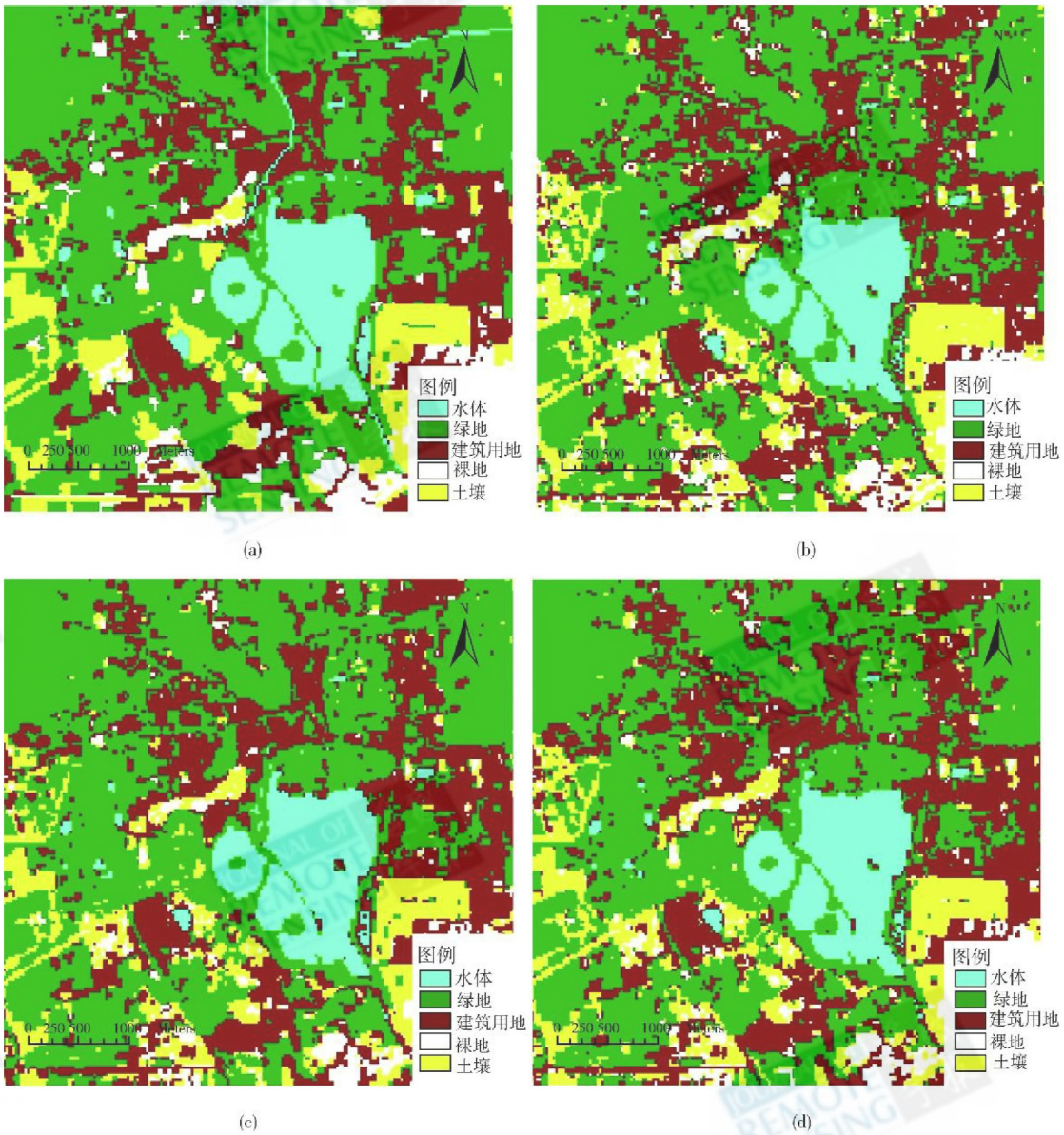

图 2　分类图像比较

(a)真实分类结果; (b)MLC 分类结果; (c) SVM 分类结果; (d)新方法分类结果

Fig. 2　Comparison of classification images by three classifiers

(a) true land use map; (b) classified image by using MLC; (c) classified image by using SVM; (d) classified image using the new method

表 1　MLC分类结果混淆矩阵

Table 1　Confusion matrix of MLC

像元数		真实类别						
		水体	绿地	建筑用地	裸地	土壤	总计	错分误差 /%
被分类别	水体	2026	6	3	0	0	2035	0. 44
	绿地	202	12247	454	65	157	13125	6. 69
	建筑用地	127	1072	6199	313	109	7820	20. 73
	裸地	0	41	389	1080	422	1932	44. 10
	土壤	0	323	298	479	2211	3311	33. 22
总计		2355	13689	7343	1937	2899		
漏分误差 /%		13. 97	10. 53	15. 58	44. 24	23. 73		
总体分类精度 84. 2%, kappa系数 0. 7675								

表 2 SVM 分类结果混淆矩阵

Table 2 Confusion matrix of SVM

像元数		真实类别						
		水体	绿地	建筑用地	裸地	土壤	总计	错分误差 /%
被分类别	水体	2092	26	2	0	0	2120	1.32
	绿地	107	12598	759	45	293	13802	8.72
	建筑用地	156	886	6228	506	299	8075	22.87
	裸地	0	10	109	884	156	1159	23.73
	土壤	0	169	245	502	2151	3067	29.87
总计		2355	13689	7343	1937	2899	28223	
漏分误差 /%		11.17	7.97	15.18	54.36	25.80		
总体分类精度 84.87%, kappa系数 0.7735								

表 3 新方法分类结果混淆矩阵

Table 3 Confusion matrix of new classifier

像元数		真实类别						
		水体	绿地	建筑用地	裸地	土壤	总计	错分误差 /%
被分类别	水体	2052	6	3	0	0	2061	0.44
	绿地	177	12510	468	48	180	13383	6.52
	建筑用地	126	955	6437	383	171	8072	20.26
	裸地	0	14	137	983	195	1329	26.03
	土壤	0	202	298	523	2353	3376	30.30
总计		2355	13689	7343	1937	2899	28223	
漏分误差 /%		12.87	8.61	12.34	49.25	18.83		
总体分类精度 86.22%, kappa系数 0.7956								

定义新方法的优势度为新方法的 kappa与两个分类器的最高 kappa之差, 两个分类器的性能差为两个分类器的 kappa之差的绝对值。图 4显示了新方法的优势度与两个分类器的性能差的关系, 可以发现, 新方法的精度提升与两个分类器的性能差成负相关, 这是因为两个分类器的性能差距越大, 错分样本集的独立性就越差, 组合的提升能力也就越有限。

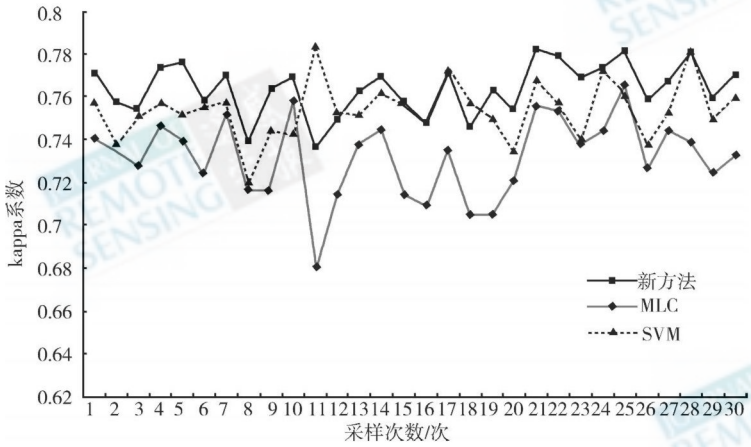

图 3 不同样本分布下 3种分类器的精度比较

Fig. 3 Comparison of classification accuracy by three classifiers using different samples

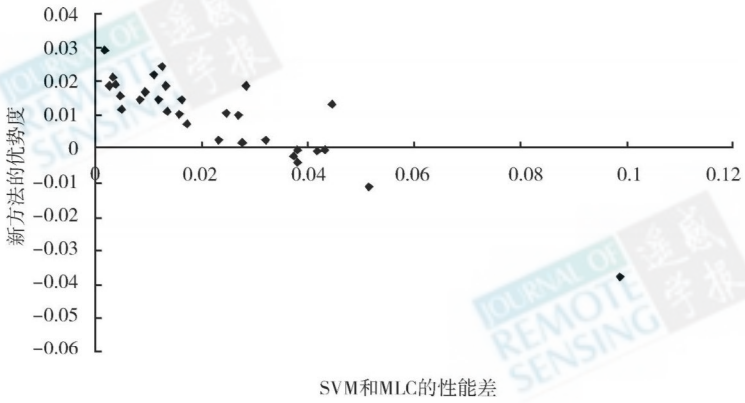

图 4　新方法优势度与 SVM、MLC的性能差的关系

Fig. 4　Relationship between the superiority of new classifier and the disparity of SVM and MLC

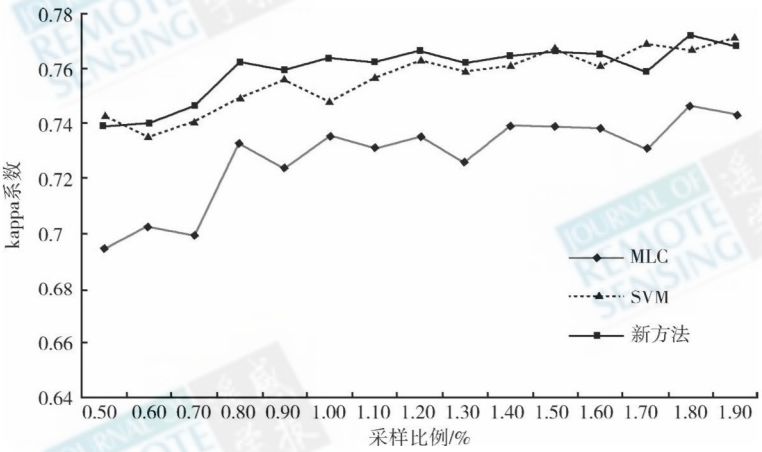

图 5　不同采样比例下 3种分类器的精度比较

Fig. 5　Comparison of classification accuracy by three classifiers using different sample amount

4.3.2　不同训练样本数量的影响

样本数量的不同也会影响分类结果, 本研究按 0. 5% —1. 9% 的比例 (样本数量从 141增加到 537)对该区域进行分层随机采样, 以考察样本数量对分类方法的影响。为消除样本位置分布不同导致的差异, 本文对每一个比例进行 10次随机采样, 再求取 kappa系数的平均值, 结果如图 5。可以看到, 在不同的样本量下, 新方法仍具有稳健的优势, 并且新方法仍然保持了 SVM 对样本数量不很敏感的特点。然而, 实验发现当样本数量增加至 1. 5% 时, 新方法不再具有优势。从实验结果看, 随着样本量的增加, MLC的分类结果并没有像理论分析一样, 有明显的提高, 这时本文的误差分析的准确性降低, 从而导致新的组合分类器未能取得好的效果。这可能是因为本文的误差分析基于总体分布是正态的假设, 但实际上这个假设很有可能并不成立, 使得对 MLC的误差估计过于乐观。

4.4　与其他组合分类器的比较

对不同分类器的规则输出进行融合, 在模式识别领域被称为测量级结合。测量级结合有很多种方式, 一般认为, 和式规则具有较强的鲁棒性[19], 因此得到广泛应用。所谓和式规则指平均各分类器的后验概率估计, 然后根据平均的后验概率, 确定待分类模式的最终类别。即

$$P_k(x)=\frac{P_{ks}(x)+P_{lm}(x)}{2} \tag{22}$$

式中, $P_{ks}(x)$, $P_{lm}(x)$和 $P_k(x)$分别表示 SVM, MLC及和式规则组合分类器的后验概率输出。取 $i=\arg\max_k[P_k(x)]$作为像元 x 的归属类别。同样地, 本文通过对该区域按 1% 分层随机采样 30次, 比较和式规则与新方法的分类精度, 图 6表示了两种方法的比较结果。可以看到, 大多数情况下, 新方法要略优于和式规则。

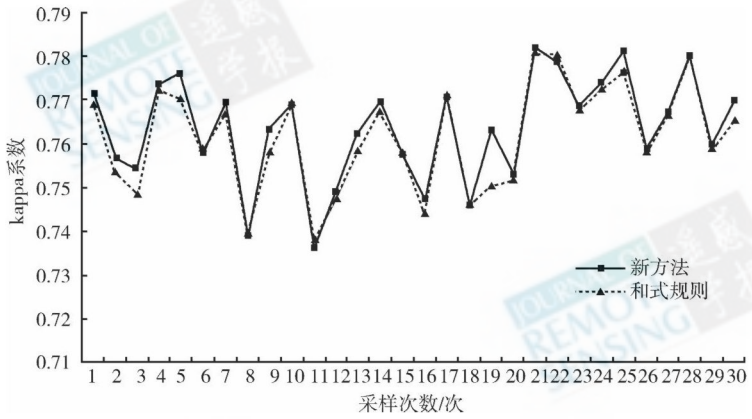

图 6 不同样本分布下两种分类器的精度比较

Fig. 6 Comparison of classification accuracy by two classifiers using different samples

5 结论和讨论

对多个分类器进行误差分析,可以有针对性的对不同数据选择不同的分类器,充分挖掘组合分类器的潜力。本文提出的误差估计方法基于经典统计理论,充分利用样本分布的信息,结合 SVM 和 MLC在每个像元上的规则输出及其误差信息,分类实验证实了该方法的优越性。分类实验通过考察在不同样本分布和样本数量下,新方法和单一分类器的分类精度比较,证实了新的组合分类器具有稳健的优势。

但是本文对分类方法的误差分析还不够全面,至少还存在以下问题需要完善:

(1)对于 MLC分类方法,只分析了样本均值的扰动对规则输出的影响,忽略了样本方差的扰动的影响;

(2)对于 SVM 分类方法,忽略了 SVM 二类分类器本身的误差;

(3)由于实际的总体也可能不服从正态分布,导致基于经典统计理论的误差分析可能会对 MLC误差的估计过于乐观。

本文尝试从理论上探讨 SVM 和 MLC的误差,给出定量表述,并以此为根据将两种分类器组合。尽管还存在诸多不足,但是显示了基于误差分析的组合分类器在遥感图像分类上的应用潜力。

参考文献 (References)

[1] Michie D, Spiegelhalter D J, Taylor C C. Machine Learning Neural and Statistical Classification[M]. Chichester: Ellis Horwood, 1994.

[2] Bo Y C, Wang J F. Combining Multiple Classifiers for Thematic Classification of Remotely Sensed Data[J]. *Journal of Remote Sensing*, 2005, **5**(9): 555—563. [柏延臣, 王劲峰. 结合多分类器的遥感数据专题分类方法研究[J]. 遥感学报, 2005, **5**(9): 555—563.]

[3] Joon Hur, Jong Woo Kim. A Hybrid Classification Method Using Error Pattern Modeling[J]. *Expert Systems with Applications*, 2008, **1**(34): 231—241.

[4] Richards J A. Remote Sensing Digital Image Analysis[M]. New York: Springer-Verlag, 1986.

[5] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

[6] Giles M Foody, Ajay Mathur, Carolina Sanchez-Hernandez, *et al*. Training Set Size Requirements for the Classification of a Specific Class[J]. *Remote Sensing of Environment*, 2006, **104**: 1—14.

[7] Qi Z Q, Tian Y J, Deng N Y. A New Support Vector Machine for Multi-class Classification[A]. Computational Intelligence and Security[C]. Berlin: Springer, 2005.

[8] Platt J C. Probabilities for Support Vector Machines[M]. Massachusetts: MIT Press, 2000.

[9] Wu T F, Lin C J. Probability Estimates for Multi-class Classification by Pairwise Coupling[J]. *Journal of Machine Learning Research*, 2004, **5**: 975—1005.

[10] BIPM-IEC-IFCC-ISO-IUPAC-IUPAP-OIML. Guide to the Expression of Uncertainty in Measurement[S]. Genevet Switzerland: ISO (International Organization for Standardization), 1995.

[11] Johnson R A, Wichern D W. Applied Multivariate Statistical Analysis[M]. New Jersey: Prentice-Hall, 1982.

[12] Hsu C W, Lin C J. A comparison of Methods for Multi-Class Support Vector Machines[J]. *IEEE Trans. Neural Networks*, 2002, **13**(2): 415—425.

[13] Duan K, Keerthi S S. Which Is the Best Multi-Class SVM Method? An Empirical Study[J]. *Multiple Classifier Systems Lecture Notes in Computer Science*, 2005, **3541**: 278—285.

[14] Lee Y, Lin Y, Wahba G. Multi-Category Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data[J]. *Journal of the*

American Statistical Association, 2004, **99** (465): 67—81.

[15] Liang S L. Quantitative Remote Sensing of Land Surfaces[M]. Jersey: A John Wiley & Sons INC., Publication, 2004.

[16] Mather P M. Computer Processing of Remotely Sensed Images (3rd ed.)[M]. Chichester: Wiley, 2004.

[17] Hixon M, Scholz D, Fuhs N. Evaluation of Several Schemes for Classification of Remotely Sensed Data[J]. *Photogrammetric Engineering and Remote Sensing*, 1980, **46**: 1547—1553.

[18] Campbell J B. Introduction to Remote Sensing (3rd ed.)[M]. London: Taylor and Francis, 2003.

[19] Tax D M J, van Breukelen M, Duin R P W, *et al*. Combining Multiple Classifiers by Averaging or Multiplying?[J]. *Pattern Recognition*, 2000, **33**: 1475—1485.

Study on Combined Classifier Based on Error Analysis

CHEN Xue-hong, CHEN Jin, YANG Wei, ZHU Kai

(State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China)

Abstract Remote sensing is widely used in mapping land use/land cover types and monitoring land use/land cover changes from regional to global scale. Supervised classification method is a powerful tool in extracting land cover and land use information from remotely sensed images. Although many supervised classification method have been developed in machine learning field, there are not a universal best performing method yet. That is, different kinds of classification methods have their own advantages and defects. This phenomenon is called selective superiority. It is necessary to explore a method that can integrate advantages of different classifiers and avoid their weakness. Combining classifiers properly may improve classification accuracy, because different classifiers may have different mistake sets. Combined classifiers have been studied widely in machine learning field; however, it was seldom studied in remote sensing image classification. This paper proposed one type of combined classifier based on error analysis, which incorporates the rule outputs of maximum likelihood classification (MLC) and support vector machine (SVM), to achieve higher classification accuracy.

MLC is the most widely used classification method in computer processing of remotely sensed images. It is based on classical statistical theory and has solid probability meanings. However, the classified accuracy of this method would be affected seriously if the training sample distribution does not follow normal distribution. SVM is a newly developed classifier, which is based on statistical learning theory. SVM is robust for small sample, and it has shown a good performance in many studies. However, the original SVM classifier is a binary classifier, which needs to be extended to a multi-class classifier through extra works. How to effectively extend binary SVM to multi-class classification is still an on-going research issue and it probably affect the performance of SVM. The new method proposed in this paper first estimates the errors of two classifiers, which are denoted by the confidence intervals of rule outputs; then combines their rule outputs with weights depending on the confidence intervals; and finally acquires a more accurate rule output. Classification experiments were conducted on case study area (Summer Palace area in Beijing). Classification accuracies of the combined classifier and two single classifiers were compared with different sample distribution and different sample amount. And the results demonstrated that the new combined classifier can acquire a higher accuracy than other two classifiers. The results also revealed that combined classifier performs better when two classifiers are more independent. Another compared experiment was done between new combined classifier and previous combined classifier by averaging, and result also showed that new method had better performance. However, there are still some defects in the new method. Firstly, error analysis is not completely finished for the two classifiers; secondly, error analysis based on classical statistical theory would be too optimistic for MLC. Although there are some disadvantages in the new combined classifier based on error analysis, it still has shown promising potential in remotely sensed image classification.

Key words combined classifier; error analysis; confidence interval; land use/land cover; remote sensing