引用格式：程希萌,沈占锋,邢廷炎,等.基于mRMR特征优选算法的多光谱遥感影像分类效率精度分析[J].地球信息科学学报,2016,18(6):815-823. [ Cheng X M, Shen Z F, Xing T Y, *et al*. 2016. Efficiency and accuracy analysis of multispectral image classification based on mRMR feature selection method. 18(6):815-823. ] **DOI:**10.3724/SP.J.1047.2016.00815

# 基于mRMR特征优选算法的多光谱遥感影像分类效率精度分析

程希萌[1,2],沈占锋[2*],邢廷炎[1],夏列钢[3],吴田军[4]

1. 中国地质大学(北京)信息工程学院,北京 100083；2. 中国科学院遥感与数字地球研究所,北京 100101；
3. 浙江工业大学计算机科学与技术学院,杭州 310023；4. 长安大学理学院,西安 710064

## Efficiency and Accuracy Analysis of Multispectral Image Classification Based on mRMR Feature Selection Method

CHENG Ximeng[1,2], SHEN Zhanfeng[2*], XING Tingyan[1], XIA Liegang[3] and WU Tianjun[4]

1. School of Information Engineering, China University of Geosciences, Beijing 100083, China; 2. Institute of Remote Sensing and Digital Earth, CAS, Beijing 100101, China; 3. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China; 4. College of Science, Chang'an University, Xi'an 710064, China

**Abstract:** Image classification is a popular research topic in the field of remote sensing. This technology has been widely used in environmental protection, military, urban planning, and other fields. Interfering by the massive feature information of remote sensing image, applying the reasonable feature selection approach in the progress of image classification becomes critical for improving the efficiency and accuracy of classification. This paper extracts the image feature data from the ZY3 satellite multispectral image of Huainan region, and studies the mRMR (minimal-Redundancy-Maximal-Relevance) feature selection method. This algorithm has a simple core principle and low requirement of data. The core problem of this algorithm is the computation of mutual information. The mRMR algorithm is initially applied in the field of bioscience, such as the gene expression analysis, and it is not widely used in the field of remote sensing. This research uses three methods (the binary discretization, histogram method and F-statistic) to realize the computation process of mRMR algorithm. And two classifiers (the C5.0 decision tree and k-nearest neighbour) are used for the classification based on three types of feature selection results and the total feature information. Moreover, the visual interpretation is used to verify the image classification results from these different methods. The study shows that the results produced by different mRMR computation processes are distinct regarding to different classifiers. In terms of efficiency, all methods can improve the efficiency of C5.0 and KNN. The classification efficiency is increased by 36.84% for C5.0 and by 72.05% for KNN. In terms of accuracy, all method can maintain the accuracy of C5.0 while improve the accuracy of KNN. The total classification accuracy and Kappa coefficient are increased for C5.0 by 0.60% and 0.80%, respectively. The total classification accuracy is increased by 4.34% and the Kappa coefficient is increased by 7.90% for KNN. In summary, the feature selection method based on the mRMR algorithm is effective in the procedure of multispectral image classification.

**Key words:** mRMR algorithm; multispectral image; mutual information; feature selection; image classification

***Corresponding author:** SHEN Zhanfeng, E-mail: shenzf@radi.ac.cn

**收稿日期:**2015-10-27;**修回日期:**2015-12-04.

**基金项目:**国家高分辨率对地观测系统重大专项(03-Y30B06-9001-13/15-01);中国科学院重点部署项目(KZZD-EW-07-02);国家高技术研究发展项目(2013AA12A401)。

**作者简介:**程希萌(1991-),男,硕士生,研究方向为空间数据挖掘。E-mail: cheng_ximeng@126.com

***通讯作者:**沈占锋(1977-),男,研究员,研究方向为遥感信息提取、分析与高性能计算。E-mail: shenzf@radi.ac.cn

**摘要**：在遥感图像分类过程中，进行合理的特征优选操作，将有助于提高分类器的分类效率及精度。本文以淮南地区资源三号卫星多光谱遥感影像数据为例，采用二值离散化、直方图法及F统计法3种计算方法实现mRMR（minimal-Redundancy-Maximal-Relevance）算法特征优选过程。根据3种方法所得到的特征优选结果及全部特征信息，分别采用C5.0决策树和K近邻2种分类器进行图像分类实验，并利用目视解译方法对不同方法组合的影像分类结果进行精度验证。实验结果表明，利用3种计算方法实现mRMR特征优选算法对不同分类器的影响程度不同：在分类效率方面，C5.0决策树分类器可提高36.84%，而K近邻分类器可提高72.05%；在分类精度方面，C5.0决策树分类器能保证分类精度大致不变，总体分类精度可提高0.60%，Kappa系数可提高0.80%，而K近邻分类器总体分类精度可提高4.34%，Kappa系数可提高7.90%。

**关键词**：mRMR算法；多光谱影像；互信息；特征优选；图像分类

## 1 引言

自20世纪后期以来，遥感技术作为一门新兴学科迅速发展[1]，遥感图像分类技术作为遥感研究的关键技术之一，已被广泛应用于环境保护[2]、军事[3]、城市规划[4]等领域。在遥感图像分类过程中，依据全部特征进行分类会对分类效率与精度造成影响，因此需要从原始特征集中选择一些关键特征，在保证不减少分类相关信息的同时减少数据总量，以达到特征优选的目的[5]。目前，较常用的特征优选方法包括主成分分析（Principle Component Analysis，PCA）、独立成分分析（Independent Component Analysis，ICA）、流形学习等。主成分分析的基本思想是搜索方向矢量，使样本之间散度最大[6]，其虽很好地考虑到特征之间的相互关系，但该方法对数据要求高，不适用于小样本的情况[7]。独立成分分析是从多元统计数据中寻找潜在成分的方法，与其他方法相比，其意义在于寻找满足统计独立和非高斯的成分，但其无法确定独立成分的顺序[8]。流形学习是一种非线性维数约简方法，能有效地探测非线性数据的内部结构，但其存在小样本问题以及噪声敏感问题[9]。这些特征优选方法均较为成熟，但也都存在一定局限性。

最小冗余最大相关（mRMR）算法[10]是一种基于互信息理论的特征优选方法，为解决这些局限性提供了契机。优选标准是使所选特征子集与类别之间的相关性最大，同时保证所选特征之间的冗余尽量小。mRMR算法原理较为简单，核心问题是对互信息的计算，其对数据要求低，具有较高的计算效率[11]，目前主要应用于医药及生命科学领域[12]。在遥感图像分类研究领域，国内已有学者开始将mRMR算法应用于遥感图像分类[13-14]，但其研究更多基于国外高光谱影像或具有高维特征信息的影像数据，对于常用国产卫星影像的研究较少。而且前人的研究更多侧重于证明mRMR算法的有效性以及将mRMR算法与其他方法相结合，并没有深入研究不同的计算方法对mRMR算法优选效果的影响，也没有深入研究mRMR特征优选对原理不同的分类器的优化效果差异。本文基于资源三号卫星多光谱数据，利用二值离散化、直方图法、F统计法3种计算方法实现mRMR优选过程，并基于优选结果利用C5.0决策树与K近邻2种分类器进行图像分类，通过对实验结果的分析，验证并比较不同方法对分类器的优化效果。

## 2 mRMR算法原理与实验流程

### 2.1 算法原理

最小冗余最大相关（mRMR）算法是由Peng在2005年提出[10]，其基本思想是利用信息论中的相关理论，以互信息量的大小作为衡量特征与特征、特征与类别标签间相关性的标准。

互信息（Mutual Information）表示2个随机变量之间的相关性[15]。基于互信息理论，Peng用特征子集中每一特征与类别标签的互信息均值表示相关性，同时用所选特征两两之间的互信息均值表示特征子集的冗余性。优选的最终目的是最大化特征子集与类别标签的相关性，同时最小化特征子集的冗余性。

将特征子集与类别标签的相关性记为 $D$，特征子集的冗余性记为 $R$。在对选出的特征子集进行评价时需同时考虑相关性与冗余性，若令相关性与冗余性具有同等权重，则特征子集的评价与选择标准如式（1）所示。

$$\max \Phi(D,R), \Phi = D - R \tag{1}$$

根据式（1）的子集评价标准，特征选择方法采用渐进式搜索算法。假设特征全集为$X$，特征子集为$S$，类别标签为$C$，$I$表示互信息。当前已进行$m-1$次选择，选出了具有$m-1$个特征的特征子集$S_{m-1}$，将要进行第$m$次选择，则选择标准如式（2）所示。

$$\max_{x_j \in X - S_{m-1}} \left[ I(x_j; C) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right] \quad (2)$$

在进行第$m$次选择时，在待选特征集$X-S_{m-1}$中对当前全部待选特征进行搜索，选出使式(2)值最大的特征$x_j$，即为第$m$次选择的特征。当进行某次选择时，若式(2)的值等于零或者小于某一设定的阈值时，则停止选择，已选特征集即为特征优选结果。搜索算法示意图如图1所示。

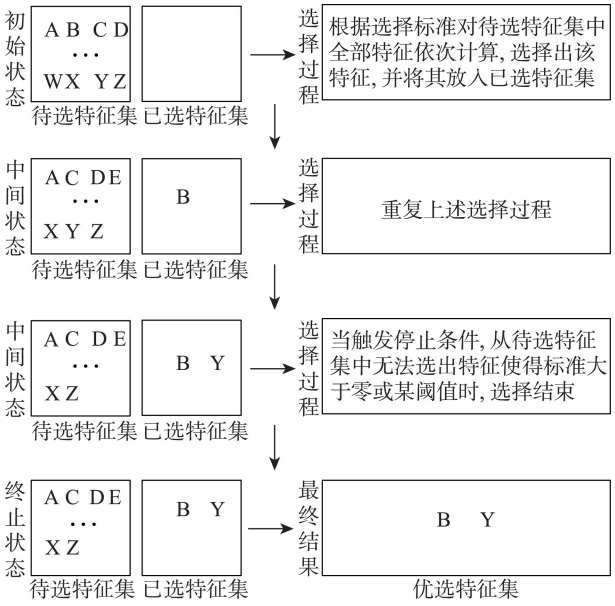



图1　渐进式搜索算法示意图

Fig. 1　Schematic diagram of the incremental search algorithm

## 2.2 实验流程

以样本特征信息以及样本地理信息为基础，将实验分为3部分，即特征优选、图像分类以及精度检验，实验流程图如图2所示。

### 2.2.1 特征优选

依据mRMR算法的计算特点以及所得特征数据的数据类型，本文分别采用二值离散化、直方图法、F统计法3种计算方法实现特征优选过程。这3种方法计算简便且理论成熟，对数据要求低，具有较强的通用性[10,12,16-17]。

(1)二值离散化

mRMR算法在计算互信息的过程中，需大量估计概率密度以及多元概率密度。针对上述问题，采用二值离散化方法(Binary Discretization，BD)[10]，将每一样本特征值按式(3)归为两类，使概率估计更容易。设特征总数为$M$，样本总数为$N$，则第$i$个样本的第$j$个特征值$x_{ij}$经过二值离散化可表示为式(3)。

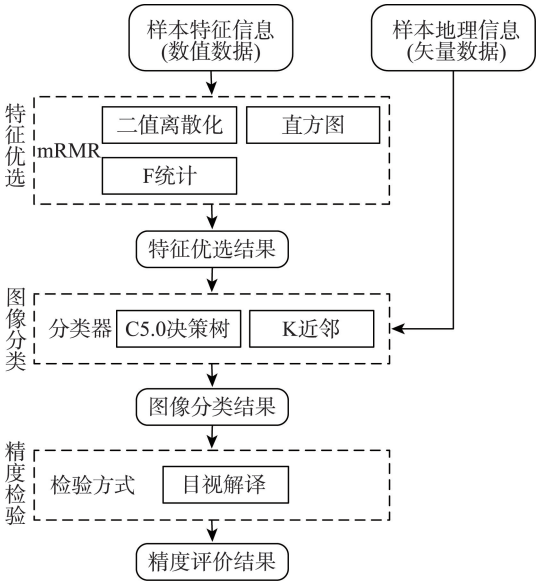



图2　实验流程图

Fig.2　Flowchart of the experiment

$$x'_{ij} = \begin{cases} 1, & x_{ij} \geqslant \bar{x}_j \\ -1, & x_{ij} < \bar{x}_j \end{cases} \quad (3)$$

式中：$x'_{ij}$表示经过二值离散化之后第$i$个样本的第$j$个特征值；$\bar{x}_j$表示所有样本的第$j$个特征值的均值。

通过二值离散化，所有样本的特征值均变为1或−1。在估计概率密度时，可以采用该值样本个数与总体样本个数的比值进行估算，而对于联合概率密度的估算也可采取同样的计算方法。

(2)直方图法

直方图法(Histogram Method，HM)是概率分布的一种估计方法，对于每一个随机变量的取值，按照一定间距取若干个间隔点，将其值域划分为若干份，通过统计每2个间隔点间的样本个数，从而绘制频率分布直方图，达到概率分布近似的目的[16]。设特征总数为$M$，样本总数为$N$，对于第$j$个特征$x_j$，确定其直方图间隔为$h_j$，分隔区间数目为$K_j$，则其近似的概率密度函数为式(4)。

$$\hat{f}(x) = \frac{1}{Nh_j} v_{kj} \left( t_{kj} < x < t_{(k+1)j} \right) \quad (4)$$

式中：$v_{kj}$表示对于第$j$个特征$x_j$，按照其直方图间隔$h_j$进行划分，落入第$k$个区间中的样本数目；第$k$个区间的左右两间隔点分别用$t_k$、$t_{k+1}$表示。其中，间隔$h$通过式(5)的经验公式[16]来确定。

$$h \approx 3.73\sigma n^{-1/3} \quad (5)$$

式中：$\sigma$为该特征样本值的标准差；$n$表示样本个数。

（3）F统计法

由于mRMR算法中互信息的计算较复杂，为此也可采用计算F统计量（F-statistic）以及皮尔逊相关系数来表示变量间的相关程度，代替互信息作为衡量相关性的标准[12]。因此，式（2）的渐进式搜索选择标准也有所变化，新的选择标准如式（6）所示。

$$\max_{x_j \in X - S_{m-1}} \left[ F(x_j, C) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} \left| r(x_j, x_i) \right| \right] \quad (6)$$

式中：$F(x_j, C)$ 表示特征 $x_j$ 与类别标签 $C$ 的F统计量；$r(x_j, x_i)$ 表示特征 $x_j$ 与特征 $x_i$ 之间的皮尔逊相关系数。

假设对于全部样本共有 $K$ 种类别标签，则F统计量的计算公式如式（7）所示[12,17]。

$$F(x_i, C) = \left[ \sum_{k=1}^{K} n_k (\bar{x}_{ik} - \bar{x}_i)^2 \Big/ (K-1) \right] \Big/ N\sigma^2 \quad (7)$$

式中：$k$ 表示不同类别标签；$\bar{x}_i$ 表示对于全部样本特征 $x_i$ 的均值；$\bar{x}_{ik}$ 表示对于类别标签为 $k$ 的全部样本特征 $x_i$ 的均值；$n_k$ 表示类别标签为 $k$ 的样本数目；$N$ 表示样本总数；$\sigma^2$ 表示合并方差。皮尔逊相关系数的计算公式如式（8）所示。

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \quad (8)$$

式中：$\bar{X}$ 与 $\bar{Y}$ 分别表示特征 $X$ 与特征 $Y$ 基于全部样本的均值；$n$ 表示样本数目。

2.2.2 图像分类

本文采用C5.0决策树及K近邻2种比较普遍，且通用性较强的分类器[18]进行遥感图像分类实验。

（1）C5.0决策树

决策树（Decision Tree）的应用范围很广，其优点在于计算复杂度低，且分类过程利于理解，缺点在于可能出现过度匹配问题[19]。本文采用C5.0决策树，它是在C4.5决策树的基础上发展而来，比C4.5决策树更加高效和稳定[18]。决策树的构建过程是以特征集中的某一特征作为根结点，依据最大信息增益的标准，选择特征集中的剩余特征进行建树并对样本集进行划分。若某划分子集中的样本类别标签均相同或某类别标签占有较大比例，则无需继续划分，该子集即为决策树的叶子结点[20]。

（2）K近邻

K近邻方法（K-Nearest Neighbour，KNN）的基本思想是对任意待测样本对象进行分类时，计算其与全部训练样本数据的欧式距离，选择距离最短的前 $K$ 个训练样本，按照这 $K$ 个训练样本的类别标签出现次数，用出现最多的类别标签标记该待测样本。K近邻方法的优点在于对分类数据的要求较低，且对异常值不敏感，缺点在于计算复杂度高、空间复杂度高[19]。本文采用的K近邻方法，固定取 $K$=5，即对于每一个待分类样本，选择与其欧式距离最短的5个训练样本数据进行分析。

2.2.3 精度检验

本文所采用的精度检验方式为目视解译，即基于图像分类结果，从全部样本中随机选取200个样本作为测试样本，通过目视解译的方式人工判别样本地物类型，并将其与分类结果进行对比，从而得出图像分类结果的精度评价。在精度检验过程中，以目视解译结果作为近似真实值难免会存在人为误差，但在难以进行实地考察以获得真实数据的前提下，目视解译作为较为准确且快速的地物类型判别方法，在遥感领域里应用广泛，且具有一定合理性。

## 3 实验结果与分析

研究区位于淮南市东北部的上窑镇。淮南市位于安徽省中北部地区（116°21′~117°11′E，32°32′~33°00′N），年平均气温为15 ℃，是中国安徽省以及华东地区重要的煤炭资源产地[21-22]。

### 3.1 数据准备

本文实验所用数据为国产资源三号卫星多光谱遥感影像，数据获取时间为2013年3月18日，影像分辨率为5.8 m。图3为经辐射校正与几何校正后的上窑镇研究区的数据影像，其覆盖范围约为62 km$^2$。

图3 研究区遥感影像图

Fig. 3 Remote sensing image of the study area

基于数据影像，利用人工勾画方式进行样本选取。由于研究区地物类型较为分明，为减少特征提取以及地物类型自身分辨难度对图像分类的影响，本文选择4种不同地物类型（建设用地、植被、水体、裸地）进行分类。影像分割方法采用均值漂移算法[23]，分割对象总数为10 449，将分割后的结果与人工选取的样本进行匹配，共计得到1591个样本对象，其中建设用地、植被、水体、裸地4种地物类型的样本个数分别为236、722、246、387。对全部的分割对象进行特征提取，为更好地检验特征优选效果，共提取特征27个，包括光谱特征、几何特征、空间关系、纹理特征4种特征类型（表1）。

**表1　特征信息**

**Tab. 1　Feature information**

| 特征类型 | 特征名称 | 特征数目 |
|---|---|---|
| 光谱特征 | 第1–4波段均值、第1–4波段标准差、NDWI均值、NDWI标准差、NDVI均值、NDVI标准差 | 12 |
| 几何特征 | 长、宽、长宽比、像元数目、边长、形状指数、角点数目 | 7 |
| 空间关系 | 紧致度、主方向 | 2 |
| 纹理特征 | 同质性、对比度、相异性、角二阶矩、熵、相关性 | 6 |

### 3.2 实验结果

经过数据准备阶段，得到研究区域的地物对象样本，并且针对每一个样本，获得了27个特征的数据信息。基于遥感特征数据，分别利用二值离散化（BD）、直方图法（HM）以及F统计法（F）3种方法实现mRMR算法的计算过程。利用3种计算方法所得mRMR算法特征优选结果如表2所示。

表2的特征优选结果显示，虽然实验中采用相同原理的mRMR算法和相同的终止条件（即若式（2）或式（6）的值为零，则优选过程结束）进行特征优选，但3组计算方法所得结果也有所不同。其中，BD与HM方法所得优选结果特征数目较接近，BD方法只比HM方法多保留了长宽比及主方向2个特征。而F方法所得结果与其他2种方法差别较大，这是由于F方法虽然在实现原理上也是基于mRMR算法，但其用F统计量与皮尔逊相关系数代替互信息作为衡量相关性的标准，在一定程度上会导致计算结果的差异。

依据上述不同的特征优选结果，在利用C5.0和KNN分类器进行图像分类时，基于每种分类器分别进行4组对比实验：基于全部特征的分类、基于BD方法特征优选结果的分类、基于HM方法特征优选结果的分类和基于F方法特征优选结果的分类。分类实验效果图中的建设用地、植被、水体、裸地地物类型分别用红色、绿色、蓝色、浅黄色表示（图4、5）。

由图4、5可发现，利用每种分类器进行4组对比实验的分类结果大体相近，参照图3的研究区影像图，分类效果图中地物类型的划分也较为清晰，这表明利用相同分类器基于mRMR特征优选结果进行分类，对分类结果影响不大，从而体现出优选过程的合理性。仔细观察分类结果还可发现，不同实验结果在部分区域也会有一定差别。例如，图4中的A、B区域表示同一区域利用不同方法得到的分类结果，可以看出二者在对建设用地及裸地地物的分类上存在明显差异，这表明特征优选也会在一定程度上影响分类器的分类结果。

### 3.3 结果分析

#### 3.3.1 C5.0决策树分类结果分析

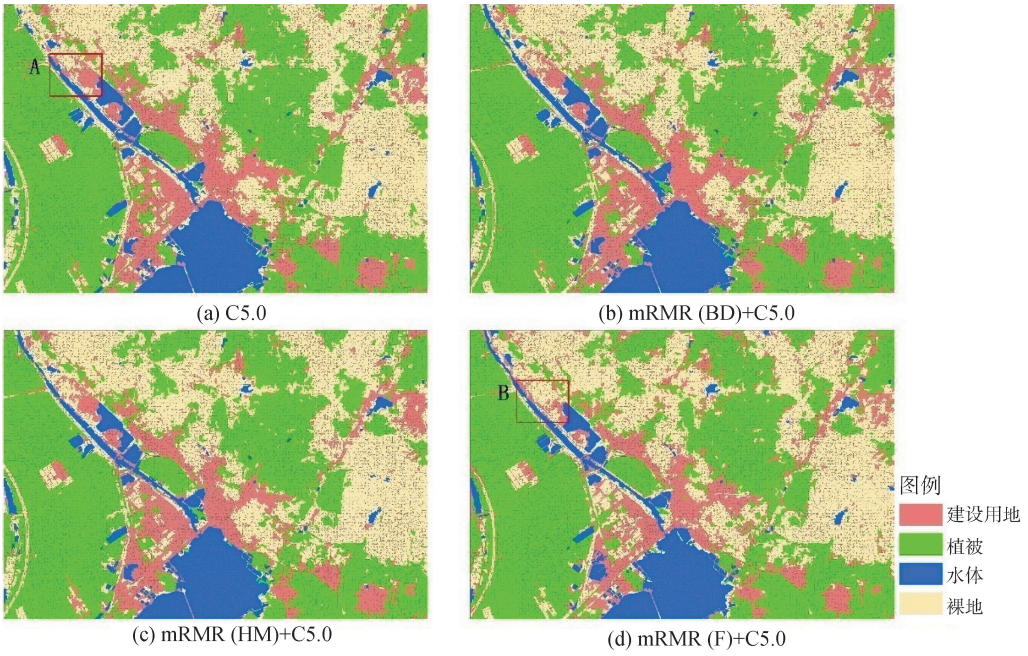
综合考虑基于C5.0分类器4组方法组合实验过程的分类效率与精度，给出对比分析结果，如表3所示。

从分类效率角度对表3进行分析。3种方法组合均可提高分类器的分类效率，其中，C5.0+F分类所用时间仅为0.852 s，与C5.0方法相比在分类效率上提高了36.84%；而C5.0+BD与C5.0+HM相比于C5.0方法在分类效率上分别提高了10.38%、14.23%。本文实验所提取的特征数目总量为27个，样本数目为10 449个。由于C5.0决策树本身是

**表2　mRMR算法特征优选结果**

**Tab. 2　The result of feature selection based on the mRMR algorithm**

| 计算方法 | 优选特征名称 | 优选特征数目 |
|---|---|---|
| BD | NDVI均值、第1波段均值、NDWI均值、第3波段均值、第4波段均值、第1波段标准差、第2波段均值、角二阶矩、相关性、同质性、第2波段标准差、第3波段标准差、相异性、对比度、熵、第4波段标准差、NDWI标准差、NDVI标准差、长宽比、主方向 | 20 |
| HM | NDWI均值、第1波段均值、相关性、同质性、第4波段均值、第1波段标准差、NDVI均值、第3波段均值、第4波段标准差、角二阶矩、第2波段均值、相异性、第3波段标准差、第2波段标准差、熵、对比度、NDVI标准差、NDWI标准差 | 18 |
| F | NDWI均值、NDVI均值、第4波段均值、第3波段均值、相关性 | 5 |

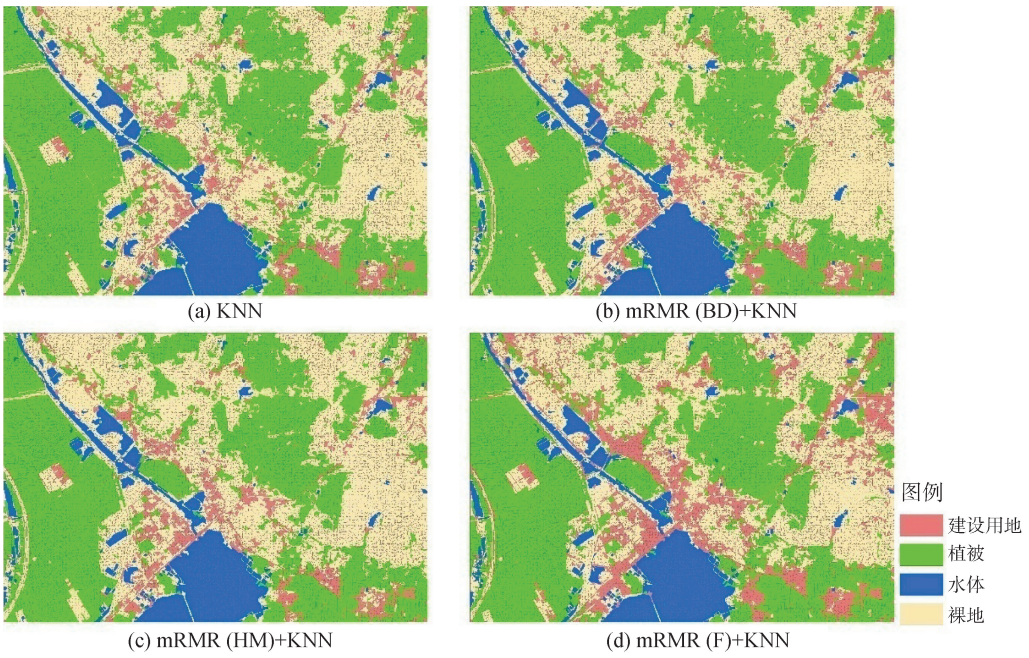



(a) C5.0 (b) mRMR (BD)+C5.0

(c) mRMR (HM)+C5.0 (d) mRMR (F)+C5.0

图4 C5.0决策树分类效果图

Fig. 4 C5.0 decision tree classification renderings




(a) KNN (b) mRMR (BD)+KNN

(c) mRMR (HM)+KNN (d) mRMR (F)+KNN

图5 K近邻分类效果图

Fig. 5 K-nearest neighbour classification renderings

一种较为高效及成熟的分类器模型，所以即便不进行特征优选而利用全部特征进行分类，分类时间也仅为1.349 s。但随着特征优选过程剔除的特征数目越多，分类时间不断减少，C5.0+BD、C5.0+HM、C5.0+F3种方法分别剔除了7、9、22个特征，其分类时间均少于利用全部特征进行分类的时间。虽然特征优选过程也需消耗一定时间，而且对于小样本数据而言，C5.0分类器具有较高的分类效率，但是特征优选所得结果可以重复利用于同组特征信息的多次研究中，通过特征优选也可得出不同地物类别与特征间的对应关系，从而在类似的研究中广泛应用。由此可见，mRMR特征优选算法可有效地提

表3　C5.0分类效率精度对比

Tab. 3　Comparison of the efficiency and accuracy between C5.0 classification methods

| 方法组合 | 分类时间/s | 总体分类精度/(%) | Kappa系数 |
|---|---|---|---|
| C5.0 | 1.349 | 83.5 | 0.747 |
| C5.0+BD | 1.209 | 82.0 | 0.724 |
| C5.0+HM | 1.157 | 84.0 | 0.753 |
| C5.0+F | 0.852 | 82.0 | 0.725 |

高C5.0的分类效率。

从分类精度对表3进行分析。3种方法组合均可保证分类精度大体不变，其中，C5.0+HM方法分类精度最高，总体分类精度与Kappa系数分别达到了84%、0.753，与C5.0方法相比精度略有提高，分别提高了0.6%及0.8%；而C5.0+BD与C5.0+F 2种方法分类精度均略低于C5.0方法。表3中数据表明总体分类精度和Kappa系数2种度量分类精度的参数呈正相关关系，这证明了实验结果的准确性。由于C5.0方法在分类过程中可通过将连续变量离散化来计算不同特征值的出现概率，从而以最大信息增益为标准选择特征进行建树，并依据决策树进行分类，这与mRMR特征优选算法在原理上存在一定相似性。并且C5.0在建树时不一定会用到全部特征，从而降低了错误特征信息对分类结果的影响，由此导致mRMR特征优选并没有显著提高C5.0的分类精度。但mRMR特征优选可以在保证精度基本不变的情况下，剔除掉一些对分类影响不大的特征，从而减少数据存储空间。

3.3.2　K近邻分类结果分析

对比分析基于KNN分类器4组方法组合实验过程的分类效率与精度，其结果如表4所示。

表4　KNN分类效率精度对比

Tab. 4　Comparison of the efficiency and accuracy between KNN classification methods

| 方法组合 | 分类时间/s | 总分类精度/(%) | Kappa系数 |
|---|---|---|---|
| KNN | 276.014 | 80.5 | 0.696 |
| KNN+BD | 212.306 | 84.0 | 0.751 |
| KNN+HM | 195.178 | 82.0 | 0.719 |
| KNN+F | 77.156 | 81.0 | 0.705 |

从分类效率对表4进行分析。3种方法组合均可以显著提高分类器的分类效率，KNN+BD、KNN+HM、KNN+F方法组合相比于KNN方法效率分别提高了23.08%、29.29%、72.05%。与C5.0分类器相比，KNN分类算法的复杂度较高，将表4统计结果与表3进行比较，可以明显反映出2种分类器在分类时间上的差异。对于本文实验而言，分类时间几乎相差100倍。本文实验的样本总数为10 449个，特征总数为27个，对于任意一个待测样本，在分类时均需计算其与已知1591个样本之间的欧式距离，计算量相对较大。考虑KNN分类器自身原理特点，利用mRMR进行特征优选以剔除冗余特征，可大大提高其分类效率。KNN+BD、KNN+HM及KNN+F 3组方法组合在特征优选过程中分别剔除了7、9、22个特征，且随着剔除特征的逐渐增多，分类效率也相应提高。在3种方法中，KNN+F方法可以将KNN分类器效率提高72.05%，在分类时间上节省近200 s，这充分证明mRMR特征优选可有效提高KNN分类效率，而对于更大的实验数据而言，特征优选过程则更为重要。

从分类精度对表4进行分析。3种方法组合均可以提高分类器的分类精度，其中KNN+BD方法拥有最高的分类精度，其在总分类精度及Kappa系数上相比于KNN方法可分别提高4.34%、7.90%。与表3实验结果类似，表4中不同实验方法的总分类精度及Kappa系数同样呈正相关，这表明了精度检验结果的可靠性。在4组实验的比较中，3组经过特征优选过程的分类结果精度均比利用KNN基于全部特征进行分类精度高，这表明对于KNN分类器而言，全部27个特征中存在一些冗余特征，从而影响分类器分类精度。KNN分类器在进行分类时需计算样本间的欧式距离，但在计算过程中并没有对参与计算的特征的重要性进行区别；而且不同特征数据单位、类型均可能不相同，虽然在计算中可以通过归一化的方式对数据进行预处理，也可通过设置权重的方式体现出不同特征的重要性差异，但如何量化权重大小也需进一步研究。所以，利用mRMR特征优选剔除在分类过程中易造成误差的冗余特征数据，会有效提高KNN分类器的分类精度。

## 4　讨论

目前，在遥感领域比较常用的特征优选方法包括主成分分析、独立成分分析与流形学习等。为比较mRMR算法与其他方法在多光谱遥感数据方面的优选效果差异，本文以主成分分析方法为例，按照相同实验流程，利用主成分分析法基于同一研究区域影像数据进行实验。

对分类结果进行检验分析，并对全部方法组合

实验结果进行统计，具体结果如表5所示。

从表5的2种优选方法的对比可看出，利用主成分分析所得结果进行影像分类，分类器效率高于mRMR算法，但其总体分类精度及Kappa系数不如mRMR算法。

**表5 基于不同特征优选算法分类效率精度对比**

**Tab. 5 Comparison of the efficiency and accuracy between different feature selection methods**

| 方法组合 | 分类时间/s | 总体分类精度/(%) | Kappa系数 |
|---|---|---|---|
| C5.0 | 1.349 | 83.5 | 0.747 |
| C5.0+mRMR | 1.073 | 82.7 | 0.734 |
| C5.0+PCA | 0.924 | 78.0 | 0.658 |
| KNN | 276.014 | 80.5 | 0.696 |
| KNN+mRMR | 161.547 | 82.3 | 0.725 |
| KNN+PCA | 88.303 | 80.0 | 0.689 |

注：C5.0+mRMR与KNN+mRMR统计数值均为二值离散化、直方图法、F统计法3种不同计算方法的统计结果平均值

mRMR算法依据不同的计算方法，优选得到的特征数目不同，所以在后续的分类过程中，不同方法对分类器效率的提高程度也有所不同。从整体来看，在效率的提高方面略逊于主成分分析，但mRMR算法考虑了特征与类别标签的相互关系，最终所优选出的特征均能很好地为分类服务，所以对分类器精度的提高程度也更明显。在数据的适应性方面，无论是连续型还是离散型数据均适用于mRMR算法，且在概率计算过程中可以弱化异常样本值，不会对整体优选结果产生很大影响。

而主成分分析则需进行数据标准化、相关系数矩阵计算、主成分函数表达式计算等过程，并根据求出的主成分函数表达式对原始数据进行处理，与mRMR算法计算过程相比，主成分分析更复杂，且对数据要求高，个别样本异常值的出现可能会造成数据标准化处理错误，从而导致整体结果出现较大误差。而且主成分分析在计算过程中并没有考虑特征与类别标签之间的关系，仅依据特征间关系划分成分，易造成关键信息丢失，从而降低分类器分类精度；但经过主成分分析后，将大量特征信息合成为主成分有效降低了特征数目，提高了分类效率。

## 5 结论

本文通过实验证明了mRMR特征优选算法在多光谱影像分类过程中的有效性。对于C5.0分类器，3种方法组合均可提高分类效率，其中C5.0+F分类效率最高，可提高36.84%；而在精度方面，3种方法组合均可保证分类精度大致不变，其中C5.0+HM具有最高的分类精度，相比于C5.0分类结果，总体分类精度提高了0.60%，Kappa系数提高了0.80%。对于KNN分类器，总体提升效果比C5.0分类器更为显著，3种方法均可提高分类效率同时提高分类精度，KNN+F可提高分类效率72.05%，而KNN+BD则可将总体分类精度和Kappa系数分别提升4.34%、7.90%。但不同方法组合对于同一分类器的提升效果也有所不同，由此则需在进行影像分类时，依据所采用的分类器以及研究目的，选择最合适的计算方法实现mRMR优选过程，本文只采用了3种计算方法，后续的研究还可以对Parzon窗[24]、Leonenko[16]等其他计算方法进行实验。在实验过程中，影像分割、特征提取、样本选择等过程均会对分类过程造成影响，为更好地验证特征优选效果，需保证在上述过程中不产生过多误差，后续的研究也可在这些方面进行深入探讨。

**参考文献(References):**

[1] 李德仁.论21世纪遥感与GIS的发展[J].武汉大学学报·信息科学版,2003,28(2):127-131. [ Li D R. Towards the development of remote sensing and GIS in the 21st century [J]. Geomatics and Information Science of Wuhan University, 2003,28(2):127-131. ]

[2] 田静,王卷乐,李一凡,等.基于决策树方法的蒙古高原土地覆盖遥感分类——以蒙古国中央省为例[J].地球信息科学学报,2014,16(3):460-469. [ Tian J, Wang J L, Li Y F, *et al*. Land cover classification in Mongolian Plateau based on decision tree method: a case study in Tov Province, Mongolia[J]. Journal of Geo- information Science, 2014,16(3):460-469. ]

[3] 许夙晖,慕晓冬,柯冰,等.基于遥感影像的军事阵地动态监测技术研究[J].遥感技术与应用,2014,29(3):511-516. [ Xu S H, Mu X D, Ke B, *et al*. Dynamic monitoring of military position based on remote sensing image[J]. Remote Sensing Technology and Application, 2014,29(3): 511-516. ]

[4] 李淑娟,王黎明,董南.城市建筑物人口时空分布模型与实验分析——以北京东华门街道为例[J].地球信息科学学报,2013,15(1):19-28. [ Li S J, Wang L M, Dong N. Simulation of urban small-area population space-time distribution based on building extraction: taking Beijing Donghuamen subdistrict as an example[J]. Journal of Geo-information Science, 2013,15(1):19-28. ]

[ 5 ] Sotoca J M, Pla F, Sánchez J S. Band selection in multi-spectral images by minimization of dependent information [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2007,37(2):258-267.

[ 6 ] 王露,龚光红.基于ReliefF+mRMR特征降维算法的多特征遥感图像分类[J].中国体视学与图像分析,2014,19(3):250-257. [ Wang L, Gong G H. Multiple features remote sensing image classification based on combining ReliefF and mRMR[J]. Chinese Journal of Stereology and Image Analysis, 2014,19(3):250-257. ]

[ 7 ] 张鹏.基于主成分分析的综合评价研究[D].南京:南京理工大学,2004. [ Zhang P. Comprehensive assessment study based on principal component analysis[D]. Nanjing: Nanjing University of Science and Technology, 2004. ]

[ 8 ] 史振威.独立成分分析的若干算法及其应用研究[D].大连:大连理工大学,2005. [ Shi Z W. Several algorithms for independent component analysis and their applications [D]. Dalian: Dalian University of Technology, 2005. ]

[ 9 ] 李波.基于流形学习的特征提取方法及其应用研究[D].合肥:中国科学技术大学,2008. [ Li B. The study of the manifold learning based feature extraction methods and their applications[D]. Hefei: University of Science and Technology of China, 2008. ]

[10] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max- dependency, max- relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(8):1226-1238.

[11] 姚旭,王晓丹,张玉玺,等.特征选择方法综述[J].控制与决策,2012,27(2):161-166, 192. [ Yao X, Wang X D, Zhang Y X, *et al*. Summary of feature selection algorithms[J]. Control and Decision, 2012,27(2):161-166,192. ]

[12] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data[J]. Journal of Bioinformatics and Computational Biology, 2005,3(2):185-205.

[13] 吴波,朱勤东,高海燕,等.面向对象影像分类中基于最大化互信息的特征选择[J].国土资源遥感, 2009,21(3):30-34. [ Wu B, Zhu Q D, Gao H Y, *et al*. Feature selection based on maxmal mutual information criterion in object-oriented classification[J]. Remote Sensing for Land & Resources, 2009,21(3):30-34. ]

[14] 邹利东,潘耀忠,朱文泉,等.结合邻域相关影像与最大相关性最小冗余性特征选择的面向对象变化检测[J].中国图象图形学报,2014,19(1):158-166. [ Zou L D, Pan Y Z, Zhu W Q, *et al*. The object- oriented change detection based on neighborhood correlation images and the minimum- redundancy- maximum- relevance feature selection [J]. Journal of Image and Graphics, 2014,19(1):158-166. ]

[15] 李亦农,李梅.信息论基础教程[M].北京:北京邮电大学出版社,2005. [ Li Y N, Li M. Information theory fundamentals[M]. Beijing: Beijing University of Posts and Telecommunications Press, 2005. ]

[16] 龚伟.基于信息熵和互信息的流域水文模型不确定性分析[D].北京:清华大学,2012. [ Gong W. Watershed model uncertainty analysis based on information entropy and mutual information[D]. Beijing: Tsinghua University, 2012. ]

[17] Dudoit S, Fridlyand J, Speed T P. Comparison of discrimination methods for the classification of tumors using gene expression data[J]. Journal of the American Statistical Association, 2002,97(457):77-87.

[18] Wu X, Kumar V, Quinlan J R, *et al*. Top 10 algorithms in data mining[J]. Knowledge and Information Systems, 2008,14(1):1-37.

[19] Harrington P.机器学习实战[M].北京:人民邮电出版社,2013. [ Harrington P. Machine learning in action[M]. Beijing: Posts & Telecom Press, 2013. ]

[20] 卢东标.基于决策树的数据挖掘算法研究与应用[D].武汉:武汉理工大学,2008. [ Lu D B. Research and application on the data mining algorithm based on decision tree [D]. Wuhan: Wuhan University of Technology, 2008. ]

[21] 杨显明,焦华富,许吉黎.基于发生学视角的淮南城市空间生长过程、特征及影响因素研究[J].地理科学,2014,34(5):563-570. [ Yang X M, Jiao H F, Xu J L. Process, characteristics and influence factors of the Huainan City's spatial expansion from the embryology perspective[J]. Scientia Geographica Sinica, 2014,34(5):563-570. ]

[22] 孙贤斌.淮南市土壤重金属污染生态研究[D].芜湖:安徽师范大学,2003. [ Sun X B. Ecological study on the soil heavy metal pollution in Huainan city[D]. Wuhu: Anhui Normal University, 2003. ]

[23] 吴田军,骆剑承,夏列钢,等.迁移学习支持下的遥感影像对象级分类样本自动选择方法[J].测绘学报,2014,43(9):908-916. [ Wu T J, Luo J C, Xia L G, *et al*. An automatic sample collection method for object- oriented classification of remotely sensed imageries based on transfer learning[J]. Acta Geodaetica et Cartographica Sinica, 2014,43(9):908-916. ]

[24] Kwak N, Choi C H. Input feature selection by mutual information based on Parzen window[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002,24(12):1667-1671.