中图法分类号: TP391.9 **文献标志码**: A **文章编号**: 1006-8961(2011)08-1474-09
论文索引信息:朱佳丽,李士进,万定生,冯钧. 基于特征选择和半监督学习的遥感图像检索[J]. 中国图象图形学报,2011,16(8):1474-1482

基于特征选择和半监督学习的遥感图像检索

朱佳丽,李士进,万定生,冯钧
(河海大学计算机与信息学院, 南京 210098)

摘 要: 随着卫星遥感技术的不断发展,基于内容的遥感图像检索技术越来越受到关注。目前该方向的研究主要集中在对遥感图像中不同特征的提取和融合方面,这些方法普遍忽略了这样一个事实:对于不同类型的检索目标,特征应该是不同的。另外,小样本问题也是遥感图像检索中一个较为突出的问题。基于以上两方面考虑,本文提出一种基于特征选择和半监督学习的遥感图像检索新方法,该方法主要包括 4 个方面:1)利用最小描述长度准则自动确定聚类数目;2)结合聚类方法和适当的聚类有效性指标选择最能表示检索目标的特征,在计算聚类有效性指数时,针对遥感图像检索特点对原有的 Davies-Bouldin 指数进行了改进;3)动态确定最优颜色特征和最优纹理特征之间的权重;4)根据最优颜色特征和最优纹理特征的权重自动确定半监督学习方法,并进行遥感图像的检索。实验结果表明,与相关反馈方法的检索效果相比,该算法在土壤侵蚀区域检索以及其他一般地表覆盖目标检索中均获得了相近的检索效果,但不需要用户多次反馈。

关键词: 基于内容的遥感图像检索;半监督学习;特征选择;聚类;聚类有效性指数

Content-based remote sensing image retrieval based on feature selection and semi-supervised learning

Zhu Jiali, Li Shijin, Wan Dingsheng, Feng Jun
(*School of Computer & Information Engineering*, *Hohai University*, *Nanjing* 210098 *China*)

Abstract: Content-based remote sensing image retrieval has been receiving more attention. Current research on content-based remote sensing image retrieval is mainly about feature extraction and feature fusion. But all these methods have ignored such a fact that retrieval objectives of different types ought to adopt different features. Besides, small size of the training set is also a challenging issue. Considering the abovementional two issues, a new remote sensing image retrieval method, which is based on feature selection and semi-supervised learning, is proposed in this paper. The new method includes four steps: 1) Determine the number of clusters automatically by using MDL criterion; 2) According to a clustering validity index, we select the optimal features which can describe the retrieval objectives best, meanwhile make some improvements in the original DB index; 3) Dynamically determine the weights of the best color feature and the best texture feature; 4) Automatically select an appropriate semi-supervised learning method and conduct image retrieval. Experiment results show that, in contrast to the retrieval performance of relevance feedback method, the method proposed in this paper obtains similar retrieval performances both in the applications of soil erosion area retrieval and general land cover retrieval in remote sensing image, as well as solving the small sample size problem.

收稿日期:2010-03-29;**修回日期**:2010-04-28
基金项目:国家自然科学基金项目(60673141)。
第一作者简介:朱佳丽(1986—),女。河海大学模式识别与智能系统专业硕士研究生,主要研究方向为基于内容的遥感图像检索与模式识别。E-mail:zhujiali123@163.com。
通讯作者:李士进,E-mail:lishijin@hhu.edu.cn。

Keywords: content-based remote sensing image retrieval; semi-supervised learning; feature selection; clustering; cluster validity index

0 引　言

随着卫星遥感技术的发展,每天可获取大量的遥感图像,如何有效地进行海量遥感图像的自动查询和检索已成为急需解决的课题。目前,国内外学者已经提出很多方法进行遥感图像的检索,如基于 Gabor 变换的纹理特征[1],颜色特征和纹理特征结合[2],纹理特征和空间信息融合[3],直方图特征相似性度量法[4],以及基于 GIS 空间语义的方法等。Zhu 等人提出利用 Gabor 纹理特征来进行航空图像的检索[1];陆丽珍等人提出融合 Gabor 纹理特征和颜色特征进行遥感图像检索,并采用纹理和颜色特征欧氏距离的线性加权来度量相似性[2];曾志明等人利用改进的共生矩阵纹理特征来进行大尺度遥感图像检索[3]。包倩和郭平等人[4]针对单波段遥感图像检索,对比研究了基于特征向量的相似性度量和基于概率的相似性度量。Ferecatu 和 Boujemaa 提出利用主动相关反馈的方法进行交互式遥感图像检索[5]。相关反馈(relevance feedback)是基于内容的图像检索(CBIR)中最常用的学习策略[5-6],它依靠人机交互过程,用户不断地进行反馈,其性能随着反馈样本集增大而提高,但同时也会大大增加用户的负担。为了减少用户多次反馈提供大量已标记样本的繁重负担,也有学者提出利用半监督学习策略进行图像检索[7-9],该策略的主要思想是利用大量的未标记示例来辅助对少量有标记示例的学习,整个学习过程不需人工干预,仅基于学习算法自身对未标记示例进行利用。

CBIR 系统主要是依靠特征提取和高维索引技术进行检索[10],其中特征的提取和使用是影响检索精度的重要因素。文献[1-4]中针对基于内容的遥感图像检索技术的研究主要集中在特征提取和融合方面,但是没有注意到这样一个事实:不同类型的检索目标,特征应该是不同的。对于同一幅图像,不同的特征在描述其内容的有效性方面也不一样,因此如果提取最能表示检索目标内容的特征应该可以有效提高检索性能。由于在基于内容的遥感图像检索中,通常只有很少的示例样本(有时甚至只有一个目标示例样本),而且要获得更多的已标记示例样本也很困难,因此采用半监督学习进行遥感图像的检索是一个较合理的选择。综合以上原因,提出一种基于特征选择和半监督学习的遥感图像检索新方法,该方法结合聚类算法和 Davies-Bouldin (DB)有效性指数[11]进行特征选择,再根据选出的最优特征构造相应的分类器进行图像检索。文献[12]提出在多分类器 Co-training 的过程中加入特征选择来进行土壤侵蚀区域的检索,取得了较好的效果。但文献[12]中特征选择过程可能会受到人为因素的干扰,且整个实验过程采用两个分类器 Co-training 的方法,忽略了不同特征之间的权重问题。而此处提出的算法是对文献[12]的延续,表现在以下几个方面:1)在聚类过程中利用最小描述长度(MDL)准则[13]来自动确定聚类数目,不但减轻了用户的负担,还可以减少人为因素对聚类结果的干扰;2)在计算聚类有效性指数时,把目标类的类内散布值直接作为总的类内散布值,把目标类与非目标类之间的类间散布值作为总的类间散布值,以突出目标类的重要性,从而有利于正确选出最能表示目标内容的最优特征;3)动态地确定最优颜色特征和最优纹理特征的二值化权重;4)根据最优特征的二值化权重来自动选择半监督学习方法;5)实验对象由原来的土壤侵蚀区域检索扩展到遥感图像中的一般地表覆盖目标检索。实验结果表明,该方法在土壤侵蚀区域检索和一般目标检索中均取得了与相关反馈方法相近的检索效果,而且还有效缩减检索时间,提高了算法效率。

1 特征选择

基于这样一种设想:不同类型的检索目标,特征应该是不同的,对于某个检索目标内容的描述,不同特征的有效性是不一样的,如果能够选择最佳表示检索目标内容的特征来进行图像检索,则可以极大地提高检索性能。本文提出在遥感图像检索中引入特征选择。

Lin 等人[14]提出颜色共生矩阵特征、基于聚类的颜色直方图、以及像素差分扫描模式 3 种特征,采用序贯前向搜索方法(sequential forward selection)进行特征选择;该方法需要在训练集上不断评估不

同特征的有效性,算法复杂度较高,不适合进行在线分析。Dy 和 Brodley 等人提出一种肺部 CT 医学图像检索的无监督特征选择方法[15],先对待检索医学图像进行分类,然后在其所在大类中进行小类细分。细分过程中他们采用基于 EM 聚类的特征选择方法,针对每次聚类结果进行鉴别分析,他们的方法计算复杂度较高,只适合离线分析。顾志伟和吴秀清等人也提出基于支持向量机以及 Adaboost 训练的医学图像检索特征选择方法[16]。Jiang 等人提出了一种在相关反馈过程中进行相似性分析的在线特征选择算法,有效提高了 CBIR 的检索性能,它是在 AdaBoost 算法的基础上发展而成的[17]。

针对遥感图像检索过程中只有一幅示例子图像的特点,通过聚类分析实现特征选择。聚类是一种典型的无监督学习技术,它根据图像内容把图像聚类到某些有意义的集合。在聚类过程中,通常由人工来确定需要预先给定的聚类数目,这不仅增加了用户的负担,而且还可能会引入人为因素对聚类结果的干扰。提出利用最小描述长度(MDL)准则[13]来自动确定聚类数目,可以有效解决这一问题。图像聚类的准则是将图像集分成多个聚类,使得位于同一聚类簇内的图像相似度尽可能大,而位于不同簇的图像相似度尽可能小。为了正确地评价聚类效果,从而客观地进行特征选择,选取合适的聚类有效性指数很重要。DB 指数是一种比较合适的聚类有效性指数,这个指数由类内散布和类间散布的比值表示,比值越小表示聚类效果越好[11]。遥感图像检索不完全是无监督的,用户最初给定的示例可以当作弱启发信息,图像特征应该有利于该图像子块和其他图像块的区别。因此我们对 DB 指数进行了一定的改进,从而更有利于特征选择,具体如下:只计算用户示例图像子块所在的目标子类的类内散布值,而不包括非目标子类的类内散布值,类间散布值也只包括非目标子类与该目标子类之间的类间散布值,而不包括非目标子类之间的类间散布值,这样不仅可以突出目标子类的重要性以及目标子类与非目标子类之间的区别,而且还可以减少计算量。

DB 指数不仅用于特征选择,还用来动态确定所选最优纹理特征和最优颜色特征的权重。对于颜色特征和纹理特征,不同值域内相同数值的 DB 指数所表示的特征差异是非等价的,因此需要通过非均匀量化来确定特征权重,这里采用二值化方法来确定特征权重。对于颜色特征,当所选最优特征的 DB 指数的倒数小于阈值 T_1 时,说明目标子类与非目标子类在颜色空间中的区别不是很明显,此时颜色特征的权重设为 0,否则为 1;对于纹理特征,当所选最优特征的 DB 指数的倒数小于阈值 T_2 时,说明目标子类与非目标子类在纹理特征空间的区别不是很明显,此时纹理特征的权重设为 0,否则为 1。计算 DB 指数时,对于纹理特征用欧氏距离来计算散布值,而对于颜色特征用直方图交距离的倒数值来计算散布值。

具体特征选择过程如下:

1)将原图像进行分块 对于遥感图像检索,合理有效的图像分块是必需的。为了避免将同一目标分入不同的小块之中,采取重叠分块策略[18]。每块大小为:$width = \min(128, \text{样本图像 } width)$,$height = \min(128, \text{样本图像 } height)$,块与块之间重叠 $width/2 \times height/2$ 像素。这样做的目的是用户在勾选查询示例图像块时,该图像块的大小代表一定的模式基元,而将子图像的长和宽限制在 128 像素内是为了避免子图像过大而导致检索结果太粗糙。块与块之间的重叠大小为 $width/2 \times height/2$ 像素,这样做是为了在检索精度和检索效率之间进行折中。众所周知,当子块以单位像素移动时,所扫描的子块几乎包括了所有待检索区域,此时检索精度是最高的,但是同时所需的时间复杂度也最高。假设一幅图像大小为 $n \times n$ 像素,模式基元的大小为 $m \times m$ ($m < 128$)像素,若按照我们的分块策略,检索的时间复杂度为 $O((n/m)^2)$,若按照单位像素移动,则检索的复杂度为 $O((n-m)^2)$。由于遥感图像的尺寸通常都比较大,而检索目标只是图像中某些分散的子块,通常 $m \ll n$,由此可见,我们的分块策略可以大大降低时间复杂度;

2)特征提取 分别提取 HSI 颜色特征、Lab 颜色特征、GLCM 纹理特征以及 Gabor 纹理特征;

3)根据 MDL 准则自动确定聚类数目 先根据最远距离准则初始化 m 个聚类中心,假设移除某个聚类中心 C_j,根据式(1)计算移除前后编码长度的总变化量,如果该变化量小于零,那么就移除这个聚类中心,否则就保留下来。依次迭代,直到没有冗余的聚类中心。最后保留下来的聚类中心数目就是自动聚类所得到的目标数目;

$$\Delta l_{C_j} = -k - n_j \log_2 p_j + \sum_{k=1,k\neq j}^{m} n_{jk} \log_2 \left(\frac{n_k + n_{jk}}{|I|} \right) + \sum_{x \in c_j} \sum_{i=1}^{d} \frac{(x_i - c_{ik})^2 - (x_i - c_{ij})^2}{2(\ln 2)\sigma^2} \tag{1}$$

式(1)中,k 表示聚类簇中心的编码长度,n_k 表示第 k 类样本的数目,n_{jk} 表示满足最近邻参考点为第 j 个聚类中心而第 2 个近邻参考点为第 k 个聚类中心的样本数目,$|I|$ 表示总的样本数目,p_j 表示第 j 类样本在总样本中所占的比重,σ 是样本数据的方差,其选取方法将在下文讨论;

4)分别在各种颜色特征空间和纹理特征空间,根据自动确定的聚类数目采用 K-means 方法进行聚类分析;

5)在各个特征空间,分别采用改进的 DB 指数进行特征有效性分析。根据式(3),分别选出 DB 指数值最小的颜色特征和纹理特征作为最优的颜色和纹理特征。

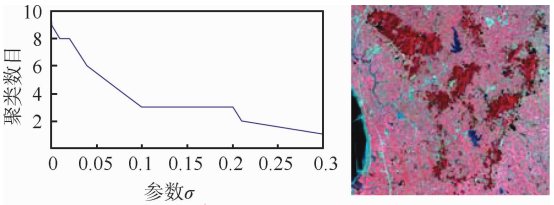
$$S_t = \frac{1}{|C_t|} \sum_{x \in C_t} D(x, p_t) \tag{2}$$

$$DB_c = \frac{1}{k-1} \sum_{\substack{i=1 \\ i \neq t}}^{k} \frac{1/S_t}{1/D(p_i, p_t)} = \frac{1}{k-1} \sum_{\substack{i=1 \\ i \neq t}}^{k} \frac{D(p_i, p_t)}{S_t}$$

$$DB_t = \frac{1}{k-1} \sum_{\substack{i=1 \\ i \neq t}}^{k} \frac{S_t}{D(p_i, p_t)} \tag{3}$$

$D(\cdot)$ 是一个距离算子,对于颜色特征,$D(\cdot)$ 表示直方图交距离;而对于纹理特征,$D(\cdot)$ 表示欧氏距离。式(2)中 t 是目标子类的簇编号,$|C_t|$ 是目标子类中的样本数目,p_t 是目标子类的聚类中心。式(3)中 k 表示总的聚类数目,p_i 表示非目标子类的聚类中心,DB_c 表示颜色特征的 DB 指数,DB_t 表示纹理特征的 DB 指数。在计算 DB_c 的时候,把交距离和的倒数作为类间散布值,而不是把交距离倒数的和作为类间散布值,这样可以减少交距离较小的单个错分样本对类间散布值的影响。

图 1 给出式(1)中参数 σ 与自动确定的聚类数目之间的关系。(b)是实验中所用的一幅原始图像,采用目测的方式可以看出,该图在颜色空间中可以聚成 3 类。从(a)可以看出,当参数 σ 的取值范围是[0.1,0.2]时,自动确定的聚类数目与理想的聚类数目相符。说明当参数 σ 的取值合适时,采用 MDL 准则自动确定聚类数目的过程是可靠的。



(a) 聚类数目与 σ 的关系　　(b) 实验所用的图像

图 1　聚类数目与参数 σ 之间的关系

Fig. 1　Number of clusters with respect to σ

2　基于半监督学习的遥感图像检索

相关反馈是 CBIR 中常用的学习策略,在每次反馈的过程中需要用户来标记正反例样本,检索性能随着标记样本数目的增加而提高,在获得较好的检索结果之前,通常需要用户进行多轮反馈操作,这大大增加了用户的负担。在基于内容的遥感图像检索中,通常只有很少的训练样本(有时甚至只有一个训练样本),而且要获得大量已标记的训练样本也很困难,因此采用半监督学习进行遥感图像检索是一个较好的选择。分别使用 Co-training 方法和自训练[19]学习方法进行图像检索。其中,Co-training 做如下假设:特征空间可以自然地分成两个,两个分类器在这两个子特征空间中进行训练。在 Co-training 的过程中,每个分类器通过添加由另一个分类器所确定的高置信度的样本来扩大自己的训练样本集。依次迭代,直到没有更多的未标记样本。文献[20]提出基于多分类器协同训练的遥感图像检索方法,采用 4 个分类器进行 Co-training 学习,取得了较好的检索效果,但时间复杂度较高。在自训练学习的过程中,先用已标记数据样本构造一个初始的分类模型,然后用这个模型去估计未标记数据的标签,用合适的选择准则选出正确的被标记数据并把它们加入到训练集中。依次迭代直到满足一定的终止条件。在 Co-training 过程中,可以通过分类器的协同性来确定迭代终止的条件。而自训练过程中只有一种特征,考虑到目标子类和非目标子类的集合都应该是以聚类中心为圆心的簇,如果某个阈值能把与目标子类最相近的非目标子类区分开,那么就可以把这个阈值设为迭代终止的条件阈值。自训练过程中阈值设定方法如图 2 所示。

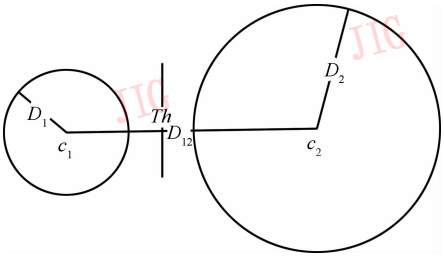


图 2 目标簇与最相邻非目标簇之间的关系

Fig. 2 Target cluster and its nearest non-target cluster

图 2 是目标簇与其最相邻的非目标簇之间的关系，Th 是所求的阈值，c_1 、c_2 分别是目标簇和最相邻非目标簇的中心，D_1 、D_2 分别是目标簇和最相邻非目标簇的半径，D_{12} 是目标簇中心和最相邻非目标簇中心之间的距离。在求取簇的半径时，一般选用簇中样本到中心的最远距离，但是考虑到簇中可能存在少量噪声样本，采用主元分析法，以距离簇中心最近的前 95% 的样本中找到的最远距离作为簇的半径。最后通过式(4)来确定阈值 Th。

$$Th = \frac{D_1}{D_1 + D_2} \times D_{12} \tag{4}$$

采用颜色和纹理这两种特征，并根据最优颜色特征和最优纹理特征的二值化权重来确定半监督学习方法。当最优颜色特征和纹理特征的二值化权重都为 1 时，采用 Co-training 方法进行检索；而当最优颜色特征和纹理特征中某一特征的权重为 0 时，采用自训练方法单独依靠权重为 1 的特征进行检索。

具体的检索算法如下：

1)特征选择　根据 MDL 准则和 DB 指数分别选出最优的颜色特征和纹理特征；

2)确定权重　根据 DB 指数分别确定最优颜色特征和纹理特征的二值化权重；

3)选择半监督学习方法　根据最优颜色特征和纹理特征的权重来选择合适的半监督学习方法；

4)图像检索　利用所选的半监督学习方法进行遥感图像检索。

3 实验结果与分析

为验证本文方法的有效性，我们对不同的地表覆盖(land cover)进行了检索实验，并把实验结果与采用相关反馈方法得到的检索结果进行比较。既有土壤侵蚀区域的检索，也有居民点、林地、湖泊围养等一般目标的检索。数据来源一部分是水利部统一提供的用于土壤侵蚀调查的数据影像文件，其数据格式为 Erdas 的 Image，数据以县为单位(部分郊区县或城关区与市辖构成一个文件)。该文件是 1999—2000 年度为主的(1∶10 0000)TM 假彩色合成数字影像，图像已经根据省级行政区域完成分景之间的镶嵌，并进行了几何纠正和统一的投影处理。卫片摄像时间为 2000 年 4—7 月。另外，我们还从 Google Earth 网站下载了太湖流域地区的一些 QuickBird 影像。具体检索过程如下：1)特征提取，实验中采用的颜色特征是 HSI 颜色特征和 Lab 颜色特征，采用的纹理特征是 Glcm 特征和 Gabor 特征；2)特征选择，分别选出最优的颜色特征和纹理特征，其中式(1)中参数 σ 的取值为 0.12；3)根据最优颜色特征和最优纹理特征的权重来确定合适的半监督学习方法，当两者的权重均为 1 时，采用 Co-training方法进行检索，当其中某个特征的权重为 0 时，采用自训练方法单独利用权重为 1 的特征所构造的分类器进行检索。在动态确定特征权重的过程中，颜色特征的阈值 $T_1 = 2.0$，纹理特征的阈值 $T_2 = 3.0$。T_1 和 T_2 是两个经验参数，后面所有不同图像实验结果的这两个阈值参数都是相同的。

图 3 是关于遥感图像中一般目标的检索结果。其中第 1 列是采用相关反馈方法分别反馈 5 次、8 次和 15 次所得的检索结果；第 2 列是加入特征选择之后，采用所选最优特征进行检索的结果；第 3 列是由专家人工给出的真正应该检索出的相似区域，其中白色细线标出的是初始检索目标示例样本，第 1 行是茅山附近林地，图像源是 TM 影像；第 2 行是苏州水乡一景，第 3 行是太湖湖西滆湖中围养，后两者都是 QuickBird 影像。图 3 中 3 幅图的颜色和纹理特征的 DB 指数的倒数均大于设定的阈值，此时颜色特征和纹理特征的权重均为 1，因此采用两个分类器 Co-training 检索方法。由图可以看出通过相关反馈方法可以获得很好的检索结果，而引入特征选择后采用两个分类器Co-training方法也能获得类似的检索效果。

图 4 是关于土壤侵蚀区域的检索结果。其中第 1 列是采用相关反馈方法分别反馈 5 次和 6 次所得的检索结果；第 2 列是引入特征选择之后，利用所选出的最优颜色特征和最优纹理特征构造的分类器进行 Co-training 检索的结果；第 3 列是由专家人工给出的真正应该检索出的相似区域，同时还给出了初

始的检索样本(用白色的细线标出)(c)是南京浦口老山附近一景,(f)是徐州附近微山湖旁一景。原图是 TM 图像,本文有缩放。由图 3 和图 4 可以看出,引入特征选择后采用两个分类器Co-training方法能获得与相关反馈类似的检索效果,而且不需要人工干预,大大减轻了用户的负担。

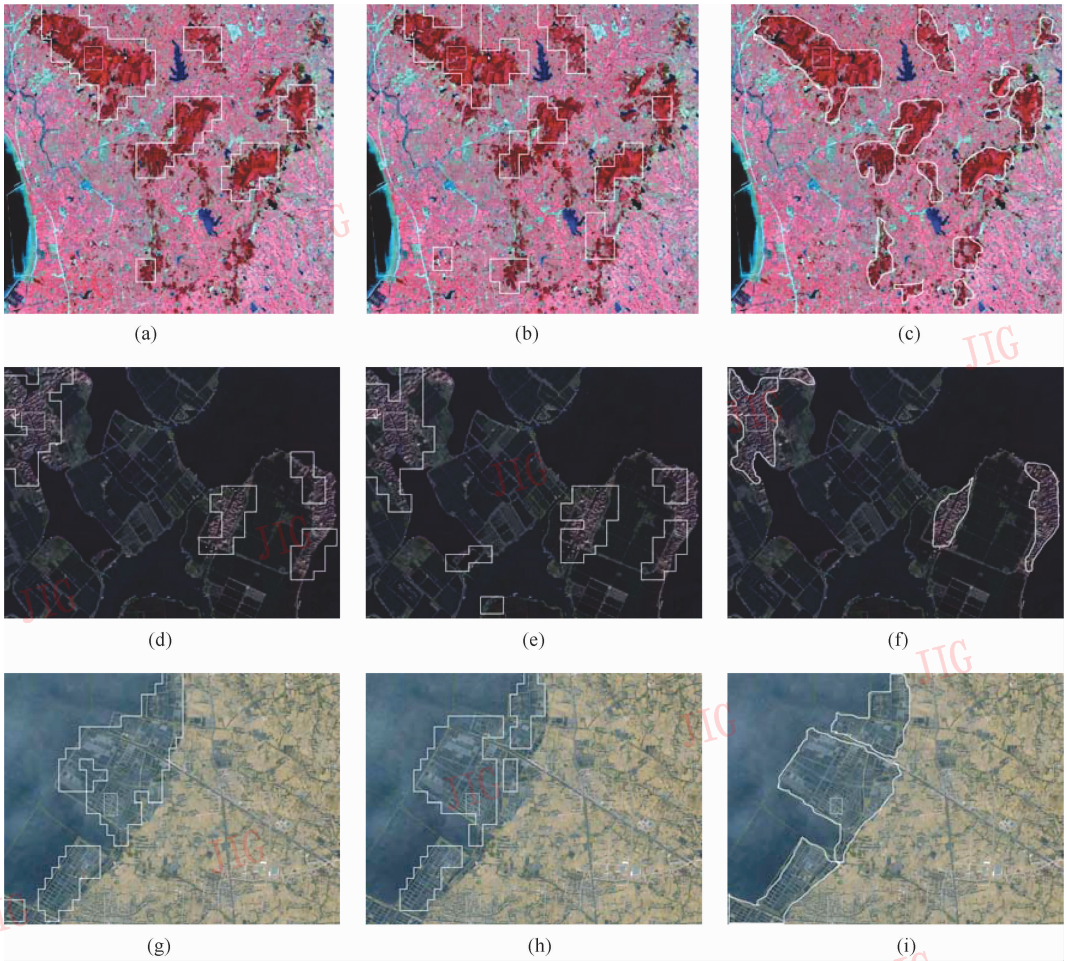


图 3　遥感图像中一般地表覆盖检索

Fig. 3　Remote sensing images retrieval of general land covers

图 5 也是关于遥感图像中一般目标的检索结果。其中第 1 列是采用相关反馈方法分别反馈 15 次和 10 次所得的检索结果;第 2 列是引入特征选择之后,采用所选出的最优特征进行自训练检索的结果;第 3 列是由专家人工给出的真正应该检索出的相似区域,其中白色细线标出的是初始的检索样本。图 5 中图像(b)最优颜色特征的 DB 指数的倒数为 1.7,小于给定的阈值 2.0,此时颜色特征的权重为 0,因此只采用最优纹理特征进行单分类器自训练检索。图像(e)的最优纹理特征的 DB 指数的倒数为 2.4,小于给定的阈值 3,此时纹理特征的权重为 0,因此采用最优颜色特征进行单分类器自训练检索。根据式(4)可知,图像(b)中颜色特征的阈值为 0.3,在自训练过程中,当不存在与示例图像之间的距离小于 0.3 的样本时,自训练迭代结束;同理,图像(e)中纹理特征的阈值为 0.05,当不存在与示例图像之间的欧氏距离小于该阈值的样本时,自训练迭代结束。由图 5 可以看出引入特征选择后采用权重为 1 的最优特征所构造的分类器进行自训练检索能获得与相关反馈类似的检索效果,而且不需要人工干预,大大减轻了用户的负担。图 5 说明当目标子类与非目标子类在某种特征空间中区别不明显时,仅利用一种区别较明显的特征进行单分类器检索就能取得较好的检索结果,且能大大减少检索时间。

图 6 分别是 4 个分类器 Co-training 检索,两个分类器 Co-training 检索以及单个分类器自训练检索

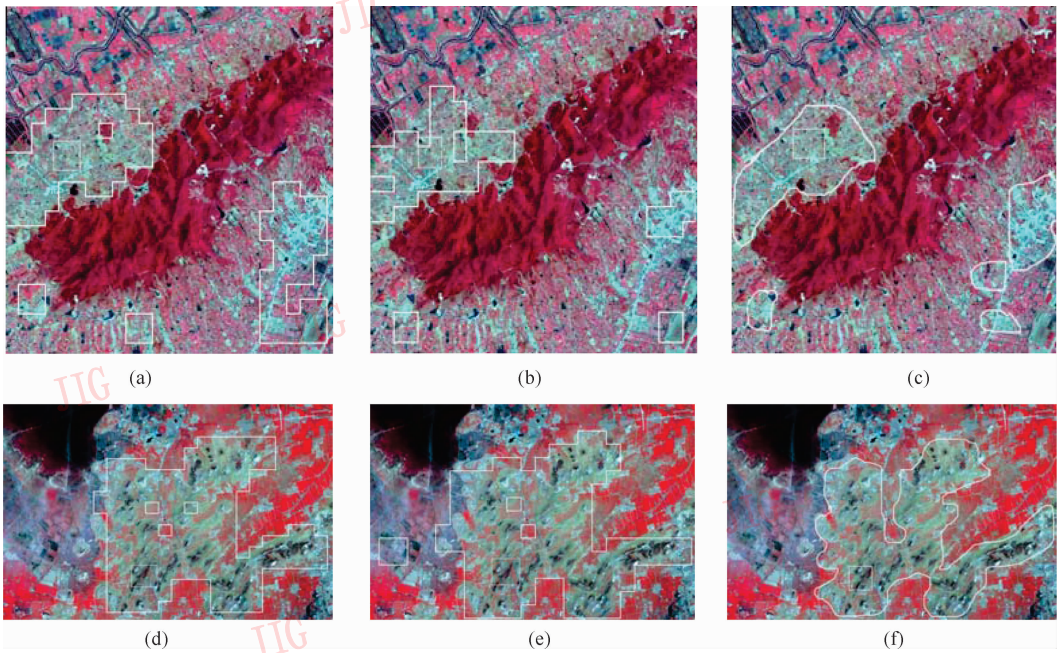


图 4 遥感图像中土壤侵蚀区域的检索结果

Fig. 4 Soil erosion areas retrieval in remote sensing images

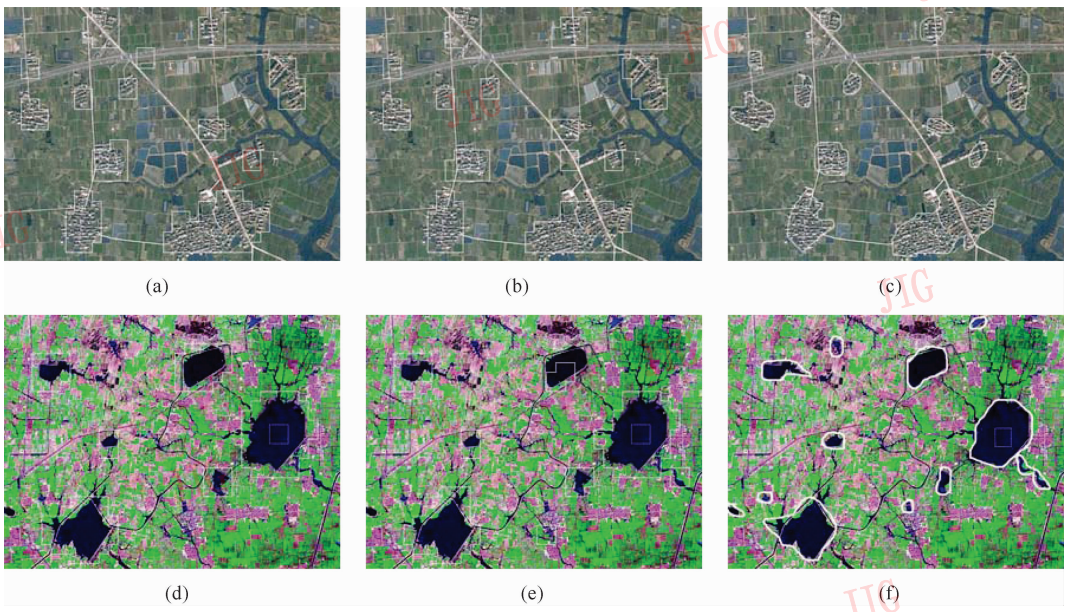


图 5 遥感图像中一般目标的检索:居民地(QuickBird 影像)和水库(TM 影像)

Fig. 5 General queries of remote sensing images retrieval: residential and reservoirs

的次数与所需时间之间的关系。该实验中采用的遥感图像集一共包含 1 025 个子图像。由图 6 可以看出,采用单个分类器检索所需的时间最少,两个分类器 Co-training 检索所需的时间次之,4 个分类器 Co-training 检索所需的时间最长。由于在 Co-training 检索过程中,需要每一个分类器对图像集中的未标记子图作出标记,然后通过融合各个分类器所做的标签来确定子图的标签,从而完成图像检索,因此对子图进行标记所需的时间会随着分类器数目的增加而增加。另外,分类器是根据未标记样本与正例样本之间的距离关系来进行标记的,随着检索次数的增加,正例训练样本的数目逐渐增多,因此图 6 中每

增加十次检索所需要的额外时间也逐渐变多。综上所述,通过引入特征选择来减少分类器的数目,可以实现检索的加速,且检索精度未见明显下降。

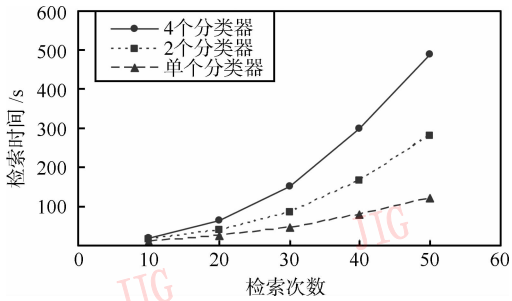


图 6　3 种不同的检索方法中,检索次数与检索时间的关系

Fig. 6　Retrieval time with respect to retrieval rounds by different retrieval methods

从实验结果可以看出,提出的基于特征选择和半监督学习遥感图像检索方法可以获得与相关反馈方法相当的检索效果,而且不需要人工干预,可以大大减轻用户的负担,同时通过特征选择减少分类器的数目,可以有效减少检索时间。这就证实了之前的设想:不同类型的图像,特征是不同的,通过特征选择选出最优的特征并结合适当的半监督学习方法进行检索,可以有效地提高检索性能。实验结果还表明,通过 DB 指数的倒数来动态确定特征的权重是可行的,文中阈值的设定也是合理的。实验结果进一步说明,在基于内容的遥感图像检索中,并不是加入的特征种类越多检索的性能就越好,反之,随着特征种类数目的增多,检索过程中需要的计算量就越大,而且还不一定能提高检索精度,因此通过特征选择选出最优特征,再利用改进 DB 指数自动确定特征的权重,并结合适当的半监督学习方法进行图像检索不仅可以提高检索精度,还能有效减少检索过程中的计算量,提高检索的速度。

4　结　论

提出一种基于特征选择和半监督学习的遥感图像检索新方法,在应用于土壤侵蚀区域的检索和遥感图像中一般目标的检索中均比未进行特征选择的检索效果要好,而且还通过 DB 指数的倒数动态地确定特征的权重,通过特征的权重来选择合适的半监督学习方法,这样不仅提高检索质量,还实现了检索的加速。但是,由于遥感图像可能存在较小目标区域的特点以及我们所采用的分块策略的局限性,对于一些分散在非目标子块中的小目标区域,还不能有效地检索。因此在下一步的研究工作中,将尝试在现有的方法中融入图像分割和基于区域的图像检索(RBIR)方法,以解决非目标子块中的小目标问题,从而进一步提高检索性能。

志谢　感谢江苏省水文水资源勘测局高祥涛高工提供本文实验遥感图像数据,并手工标注检索目标区域和 ground-truth 结果。

参考文献(References)

[1] Zhu Bin, Marshall R, Hsinchun C. Creating a large-scale content-based airphoto image digital library[J]. IEEE Trans. on Image Processing, 2000, 9(1): 163-167.

[2] Lu Lizhen, Liu Renyi, Liu Nan. Remote sensing image retrieval using color and texture fused features[J]. Journal of Image and Graphics, 2004, 9(3): 328-332. [陆丽珍,刘仁义,刘南. 一种融合颜色和纹理特征的遥感图像检索方法[J]. 中国图象图形学报,2004,9(3):328-332.]

[3] Zeng Zhiming, Li Feng, Fu Kun, et al. A method for extracting texture characters from large-scale remote sensing image[J]. Journal of Wuhan University(Information and Science edition), 2005, 30(12): 1080-1083. [曾志明,李峰,傅琨,等. 一种大尺寸遥感图像基于内容检索的纹理特征提取算法[J]. 武汉大学学报:信息科学版,2005,30(12):1080-1083.]

[4] Bao Qian, Guo Ping. Comparative studies on similarity measures for remote sensing image retrieval based on histogram[J]. Journal of Remote Sensing, 2006, 10(6):893-900. [包倩,郭平. 基于直方图的遥感图像相似性检索方法比较[J]. 遥感学报,2006, 10(6): 893-900.]

[5] Marin F, Nozha B. Interactive remote sensing image retrieval using active relevance feedback[J]. IEEE Transactions on Geoscience and Remote Sensing, 2007, 45(4): 818-826.

[6] Rui Y, Huang T S, Ortega M, et al. Relevance feedback: A power tool for interactive content-based image retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1998,8(5): 644-655.

[7] Yao Jian, Zhang Zhongfei, Antani S, et al. Automatic medical image annotation and retrieval using SEMI-SECC[C]// Proceedings of IEEE International Conference on Multimedia and Expo. Piscataway, NJ, United States: IEEE Press, 2006: 2005-2008.

[8] Dong Anlei, Bir B. A new semi-supervised EM algorithm for image retrieval[C]// Proceeding of IEEE Computer Society Conference on Computer Vision and Patten Recognition. Piscataway, NJ, United States: IEEE Press, 2003: 110-116.

[9] Sharma A, Hua Gang, Liu Zicheng, et al. Meta-tag propagation by co-training an ensemble classifier for improving image search relevance [C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ, United States: IEEE Press, 2008: 1-6.

[10] Li Zhixin, Shi Zhiping, Li ZhiQing, et al. A survey of semantic mapping in image retrieval [J]. Journal of Computer-Aided Design & Computer Graphics, 2008, 20(8): 1085-1096. [李志欣,施智平,李志清,等. 图像检索中语义映射方法综述[J]. 计算机辅助设计与图形学学报, 2008, 20(8): 1085-1096.]

[11] Davies D L, Bouldin D W. A cluster separation measure [J]. IEEE Trans. Pattern Anal. Machine Intell, 1979, 1(4): 224-227.

[12] Li Shijin, Zhu Jiali, Gao Xiangtao, et al. Soil erosion remote sensing image retrieval based on improved semi-supervised learning[C]// Proceedings of the 2009 Chinese Conference on Pattern Recognition. Piscataway, NJ, United States: IEEE Press, 2009: 395-399.

[13] Horst B, Ales L, Alexander S. MDL principle for robust vector quantization [J]. Pattern Analysis & Applications, 1999, 2: 59-72.

[14] Lin C, Chen R, Chan Y, A smart content-based image retrieval system based on color and texture feature[J]. Image and Vision Computing, 2009, 27(6): 658-665.

[15] Dy J, Brodley C, Kak A, et al, Unsupervised feature selection applied to content-based retrieval of lung images [J]. IEEE Trans. PAMI, 2003, 25(3): 373-378.

[16] Gu Zhiwei, Wu Xiuqing, Jing Hao, et al. A feature selection based approach in medical image retrieval[J]. Chinese Journal of Biomedical Engineering, 2007, 26(01): 30-34. [顾志伟, 吴秀清, 荆浩,等,一种基于特征选择的医学图像检索方法[J]. 中国生物医学工程学报, 2007, 26(01): 30-34.]

[17] Jiang Wei, Er Guihua, Dai Qionghai, et al. Similarity-based online feature selection in content-based image retrieval [J]. IEEE Transactions on Image Processing, 2006, 15(3): 702-712.

[18] Li Deren, Ning Xiaogang. A new image decomposition method for content-based remote sensing image retrieval[J]. Transactions on Geomatics and Information Science of Wuhan University, 2006, 31(8): 659-662. [李德仁,宁晓刚. 一种新的基于内容遥感图像检索的图像分块策略[J]. 武汉大学学报:信息科学版, 2006, 31(8): 659-662.]

[19] Luca D, Fabio R. Using co-training and self-training in semi-supervised multiple classifier systems [C]// SSPR&SPR 2006, LNCS 4109. Berlin Heidelberg: Springer-Verlag, 2006: 522-530.

[20] Li Shijin, Tao Jian, Wan Dingsheng, et al. Remote sensing image retrieval based multiple classifiers Co-training[J]. Journal of Remote Sensing, 2010, 14(3): 493-506. [李士进,陶剑,万定生,等. 多分类器协同训练的遥感图像检索[J]. 遥感学报, 2010, 14(3): 493-506.]