文章编号:1002-3682(2015)01-0047-08

基于决策树算法的 Landsat 8 OLI 影像海岸类型自动识别方法*

王常颖[1],谭　萌[2]

(1.青岛大学 信息工程学院,山东 青岛 266071;2.国家海洋局 北海信息中心,山东 青岛 266033)

摘　要:针对基岩、人工、砂质和淤泥质海岸四种海岸类型,采用数据挖掘算法中的 C4.5 决策树算法对 Landsat 8 OLI 多光谱影像数据进行分析。首先得出 4 种海岸类型中所包含的海水、陆地、植被、养殖区、淤泥、沙滩六种地物的识别规则;然后利用不同海岸类型所包含的地物类型差异,分别提出了基岩、人工建筑、人工养殖、沙滩、淤泥五种海岸类型的自动识别规则。实例验证结果表明,对于地物特征明显的人工海岸和基岩海岸,我们提出的海岸类型自动识别方法的精度可达到 100%,而对于光谱相近的淤泥质海岸和人工养殖海岸其识别精度较低,大于 80%。

关键词:Landsat 8;海岸类型;自动识别;数据挖掘;C4.5 决策树算法

中图分类号:TP79　　**文献标识码**:A

海岸的演变受多种因素影响,不同海岸类型的演变特征不同,正确划分海岸类型有利于找出海岸演变规律,从而预测未来,为合理开发利用海岸提供指导。由于影响海岸演变的因素十分复杂,分类方法不统一导致海岸分类也不统一。我们采用我国近海海洋综合调查与评价专项中海岛海岸带遥感调查的分类体系,将海岸类型分为基岩海岸、人工海岸、淤泥质海岸、砂质海岸和生物海岸五大海岸类型[1]。

到目前为止,对于海岸带区域,人们较多关注海岸线提取技术[2-6]和海岸带地物分类技术[7]的研究,对于海岸类型自动识别技术[8]关注较少。本文以基岩海岸、砂质海岸、淤泥质海岸和人工海岸为研究对象,探讨这 4 种海岸类型的自动识别方法。

1　Landsat 8 OLI 数据与海岸类型特征

1.1　Landsat 8 OLI 影像

Landsat 8 卫星于 2013-02-12 在美国加利福尼亚州的范登堡空军基地(VandenbergAir

* **收稿日期**:2014-10-23

资助项目:山东省科技发展计划——基于数据挖掘的 TM 影像山东省海岸类型自动识别及海岸线快速提取技术研究(2011YD15005);国家自然科学基金——海洋灾害大数据分析的系统模型研究及应用(41476101);山东省自然科学基金重点项目——复杂海量数据可计算建模及基于数据挖掘的机理发现方法的研究(ZR2012FZ003);青岛市科技计划——可计算复杂网络模型研究及其在海洋生态灾害预警预测与海洋执法中的应用(13-1-4-121-jch)

作者简介:王常颖(1980-),女,讲师,博士,主要从事海洋遥感与数据挖掘方面研究. E-mail:wcing80@126.com

(陈　靖　编辑)

Force Base,California)发射成功[9]。Landsat 8 携带了 2 个主要载荷：运行陆地成像仪(Operational Land Imager,OLI) 和热红外传感器(Thermal Infrared Sensor,TIRS)。其中,OLI 的波段设置见表 1。

表 1 Landsat 8 OLI 影像波段设置

Table 1 Band set of Landsat 8 OLI image

波 段	波长/μm	地面采样距离/m
1	0.433～0.453	30
2	0.450～0.515	30
3	0.525～0.600	30
4	0.630～0.680	30
5	0.845～0.885	30
6	1.560～1.660	30
7	2.100～2.300	30
8(全色)	0.500～0.680	15
9	1.360～1.390	30

1.2 海岸类型特征

基岩海岸是海浪长期侵蚀海岸边的岬角所形成的,其解译特征是海岬角以及直立陡崖的水陆直接相接地带,主要由植被、海水和陆地建筑物三种地物类型组成。人工海岸包括围堤、码头、防波堤、护岸(坡)等多种用途的海岸人工构筑物,人工海岸的地物主要以陆地建筑、海水和养殖区为主。砂质海岸主要由波浪塑造而成,其海滩物质组成主要为砂,少量砾石组成的海滩也包括在内。淤泥质海岸是主要由潮汐作用塑造的低平海岸,滩面坡度平缓,滩面宽度可达数公里甚至更宽,一般有潮沟发育,一些潮沟上端与河流入海口相接。这种海岸受上冲流的影响,在潮间带之上向陆一侧有一条耐盐植物生长状况明显变化的界线。

由 4 种海岸类型的特征分析可知,分布在海岸带区域的地物类型主要是陆地、沙滩、海水、植被、养殖区和淤泥。因此,要想实现海岸类型的自动识别,关键是要将不同海岸类型中所包含的地物类型进行准确识别。

2 基于决策树算法的海岸带地物类型分类规则分析

由以上各种海岸类型的分析可知,海岸带区域的地物类型主要为陆地、沙滩、海水、植被、养殖区和淤泥。本文采用数据挖掘算法中的 C4.5 决策树[10]算法进行地物识别规则的分析。

2.1 特征分析

为了能够得出高精度的地物识别规则,特征选择是关键,其原则是尽量选择能够区分各类地物的特征。因此,本文以 2013-04-21 获取的青岛附近海域的 Landsat 8 OLI 影像为数据源,分别选择陆地建筑(简称为陆地)、沙滩、海水、植被、养殖区(包括浑浊海水)和淤泥样本进行光谱比较,如图 1 所示。由于第 8 波段为全色波段,分辨率为 15 m,所以本文只对除波段 8 以外的其他波段进行比较。

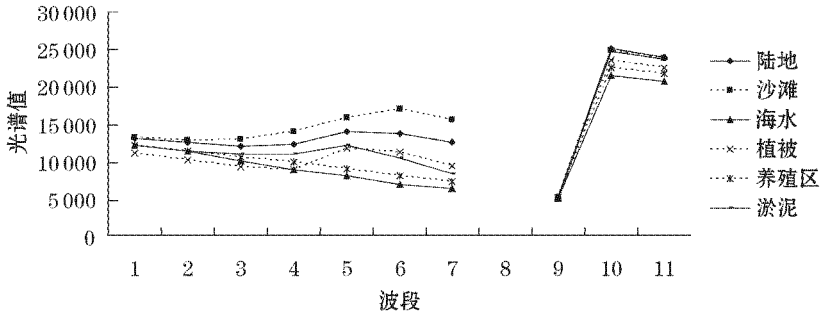

图 1　海岸带地物光谱比较分析

Fig. 1　Comparative analysis of the spectra of the coastal land features

根据光谱比较，发现各地物的光谱值中，波段 4～波段 7 具有比较明显的差异，而波段 1～波段 3、波段 10 和波段 11 中对各类地物的区分性不强。对于海水和植被样本，波段 2 略高于波段 3 的光谱值，而其他地物则刚好相反；沙滩和陆地样本的波段 4 略高于波段 3 的光谱值，而其他地物相反；养殖区和海水样本的波段 5 略高于波段 6 的光谱值，而其他地物相反；沙滩样本的波段 6 高于波段 5，而其他地物则相反。因此，为了能得出最优的分类规则，本文选择波段 4(b4)、波段 5(b5)、波段 6(b6)、波段 7(b7)、波段 2-波段 3(b2-b3)、波段 3-波段 4(b3-b4)、波段 4-波段 5(b4-b5)、波段 5-波段 6(b5-b6)、波段 4-波段 6(b4-b6)等特征，进行分析前的数据准备。

2.2　规则挖掘算法

将准备好的陆地、海水、沙滩、植被、养殖区和淤泥等地物样本的 b4，b5，b6，b7，b2-b3，b3-b4，b4-b5，b5-b6 和 b4-b6 共 9 个特征值作为待挖掘数据集 T，采用 C4.5 决策树方法进行地物分类规则挖掘。本文的类别空间设为{陆地，海水，沙滩，植被，养殖区，淤泥}，准备的数据描述特征共 9 个属性。

对于每个非类别的连续属性 D(可为 9 个特征值中的任何一个)，其取值范围内的任何一个值均可视为一个分割点，能够将数据集 T 分成集合 T_1 和 T_2，其中 T_1 为大于分割点的数据集，T_2 为小于分割点的数据集。为了得到最优分割点，使得被分割后 T_1 和 T_2 数据集中样本的类别尽可能一样，分别计算所有分割点的增益率，选择增益率最大的分割点作为该属性的划分(最优分割点)。增益率(GainRatio)计算公式如下：

$$GainRatio(D,T)=\frac{Gain(D,T)}{SplitInfo(D,T)} \tag{1}$$

$$SplitInfo(D,T)=I(|T_1|/|T|,|T_2|/|T|) \tag{2}$$

$$Gain(D,T)=Info(T)-Info(D,T) \tag{3}$$

$$Info(D,T)=\sum_{i=1}^{2}|T_i|/|T|\times Info(T_i) \tag{4}$$

式中，$|T|$表示数据集 T 中的样本数量。

对于类别空间{陆地，海水，沙滩，植被，养殖区，淤泥}，若给定数据集 T 的概率分布 $p=(p_1,p_2,p_3,p_4,p_5,p_6)$(其中 $p_i(i=1,2,\cdots,6)$为数据集 T 中第 i 种地物所占的比例)，则由该分布传递的信息量称为 P 的熵：

$$Info(T) = Info(P) = -(p_1 \times \log_2(p_1) + p_2 \times \log_2(p_2) + \cdots + p_6 \times \log_2(p_6)) \quad (5)$$

上面的算法描述了对于一个属性 D，通过增益率最大选择出最优的分割点，实现连续属性划分的过程。一旦找到 9 个属性的最优分割点后，那么下一步建立分类决策树时，则选择 9 个属性中增益率最大的属性来建立树节点，然后再对分割后的子数据集 T_1 和 T_2 寻找最优树节点，继续建树，直到每个子数据集中的样本均是同样类别为止。因此，本文建立的决策树如图 2 所示。

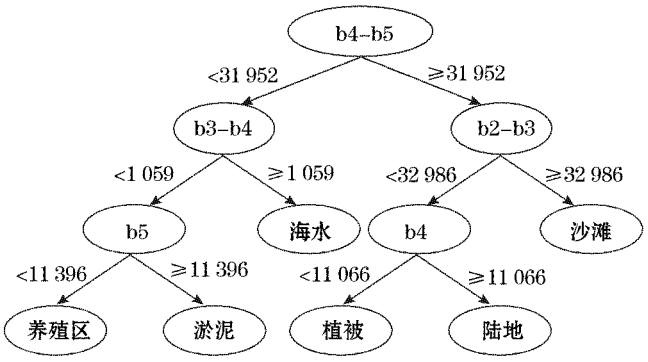

图 2 数据挖掘得出的分类决策树

Fig. 2 The mined classifying decision tree

由图 2 的分类决策树可以得出如下海水、淤泥、养殖区、沙滩、植被和陆地的分类识别规则。

规则 1：如果 b4-b5＜31 952，b3-b4＜1 059 且 b5＜11 396，则类别为“养殖区”；

规则 2：如果 b4-b5＜31 952 且 b3-b4≥1 059，则类别为“海水”；

规则 3：如果 b4-b5＜31 952，b3-b4＜1 059 且 b5≥11 396，则类别为“淤泥”；

规则 4：如果 31 952≤b4-b5＜64 664 且 b2-b3≥32 986，则类别为“沙滩”；

规则 5：如果 31 952≤b4-b5＜64 664，b2-b3＜32 986 且 b4＜11 066，则类别为“植被”；

规则 6：如果 31 952≤b4-b5＜64 664，b2-b3＜32 986 且 b4≥11 066，则类别为“陆地”。

2.3 规则验证

在 2013-04-21 获取的青岛附近海域 Landsat 8 影像中截取 2 个小的研究区域作为验证对象，如图 3a 和 4a 所示。采用本文的地物分类规则，进行 6 种地物的识别，识别结果如图 3b 和 4b 所示。本文得出的海岸带地物分类规则能够较高精度的识别出海水、陆地、植被、沙滩、淤泥、养殖区-浑水六种地物类型。由于本文提出的方法中没有涉及噪声后处理步骤，因此会有部分区域误检测，但对于本文下一步的海岸类型自动识别没有太大的影响，这些误检测区域可以忽略。

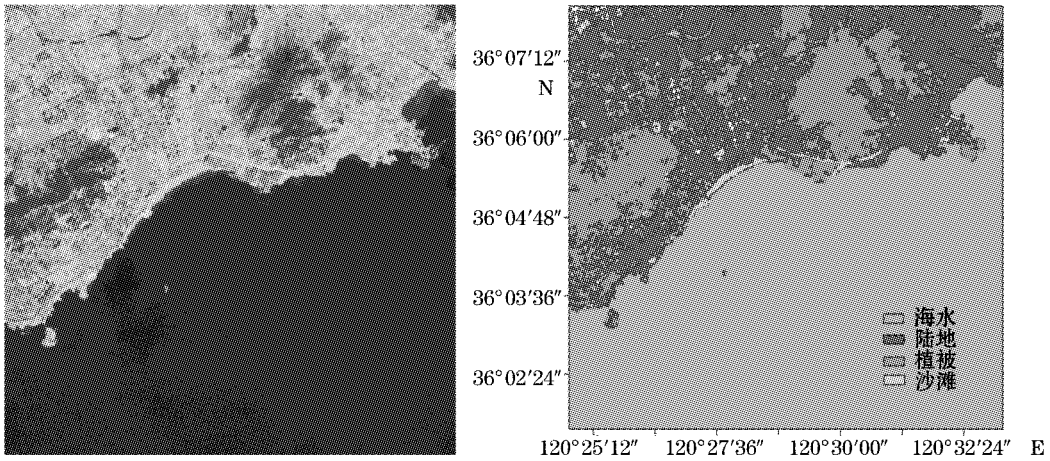

图 3　基于规则的青岛附近海岸带区域图

Fig. 3　The image of the coastal area nearby Qingdao based on the rules

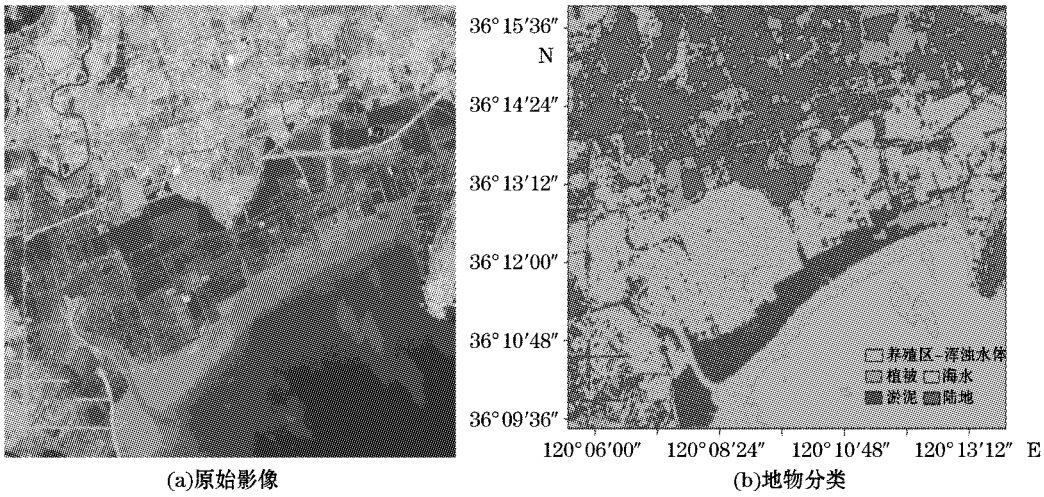

图 4　基于规则的胶州湾部分海岸带图

Fig. 4　The image of parts of the coast around the Jiaozhou Bay based on the rules

3　海岸类型自动识别方法

3.1　海岸类型中各地物所占比例特征分析

因砂质海岸中沙滩所占的比例较小，而淤泥质海岸中淤泥所占比例较大，本文进行海岸带地物分析时窗口选为 80×80，大致能够体现各种海岸类型的特点(图 5)。其中，根据人工建筑物的不同，人工海岸将分为人工建筑海岸和人工养殖海岸两种类型分别讨论。

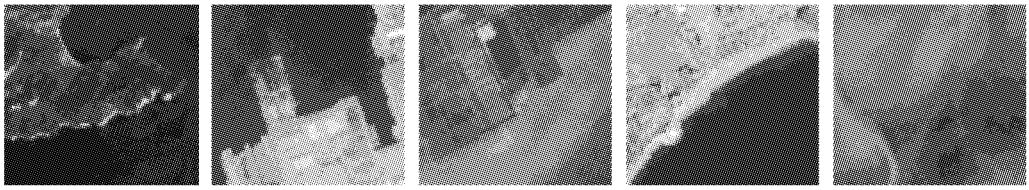

图 5 不同海岸类型比较

Fig. 5 Comparison among different types of coasts

针对基岩海岸、人工养殖区海岸、人工建筑海岸、砂质海岸和淤泥质海岸五种海岸类型，选取样本分析不同海岸类型区域包含的海岸带地物类型比例之间的差异，如图 6 所示。

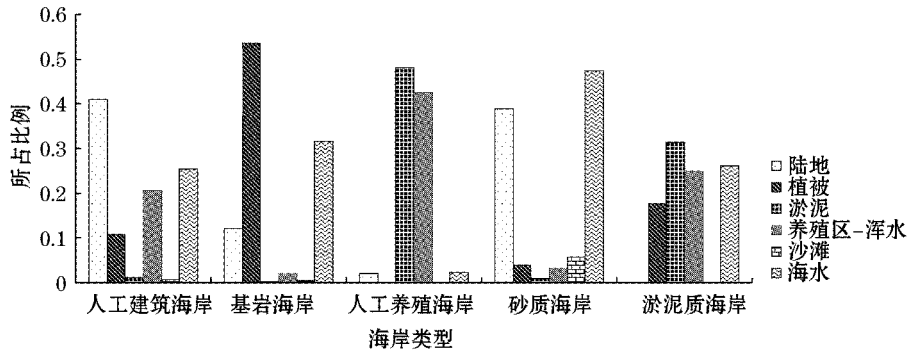

图 6 不同海岸类型中地物所占比例分析比较

Fig. 6 Analysis and comparison of the proportion of the land feature types in different types of coasts

3.2 海岸类型自动识别方法

基于不同海岸类型中所包含地物所占比例的不同，本文提出下面的海岸类型识别规则：

规则 1：如果最优三类地物为陆地、海水和养殖区-浑水，则为人工建筑海岸；

规则 2：如果最优三类地物为植被、海水和陆地，则为基岩海岸；

规则 3：如果最优三类地物为淤泥、养殖区-浑水和海水，且陆地建筑所占比例大于植被，则为人工养殖海岸；

规则 4：如果最优三类地物为海水、陆地和沙滩，则为砂质海岸；

规则 5：如果最优三类地物为淤泥、海水和养殖区-浑水，植被所占比例大于陆地，则为淤泥质海岸。

为验证本文提出的海岸类型识别规则的有效性，选择窗口为 80×80 大小的基岩海岸、人工建筑海岸、人工养殖海岸、砂质海岸和淤泥质海岸验证实例各 10 景，采用本文提出的海岸类型识别规则进行海岸类型自动识别，实验结果如表 2 所示。

表 2　实验结果及精度

Table 2　The experimental results and the accuracy

实际类型	识别类型					
	基岩海岸	人工建筑海岸	人工养殖海岸	砂质海岸	淤泥质海岸	识别精度
基岩海岸	10	0	0	0	0	100%
人工建筑海岸	0	10	0	0	0	100%
人工养殖海岸	0	0	8	0	2	80%
砂质海岸	1	1	0	8	0	80%
淤泥质海岸	0	0	3	0	7	70%
平均识别精度						86%

注：空白处为无数据

由表 2 可知，对于基岩海岸和人工建筑海岸的识别精度为 100%；人工养殖海岸中养殖区与淤泥的光谱差异较小，因此存在被误识别为淤泥质海岸的区域，但识别精度也可达到 80%；砂质海岸中的沙滩面积的大小直接影响到砂质海岸的识别精度，因此对于沙滩面积较小的砂质海岸区域，容易误识别为基岩海岸和人工建筑海岸，但识别精度也可达到 80%；而淤泥质海岸中的淤泥与养殖区较容易混淆，所以有可能误识别为人工养殖海岸，但识别精度也可达到 70%；总体上看，本文提出的海岸类型自动识别方法能够较高精度的识别出不同的海岸类型，总体精度将达到 80%以上。

4　结论与讨论

针对我国基岩、人工、砂质和淤泥质四大主要海岸类型，以 Landsat 8 OLI 多光谱影像为数据源，采用数据挖掘算法中的 C4.5 决策树算法，首先挖掘得出了这四大海岸类型中所包含的海水、陆地、养殖区、淤泥、沙滩等地物的识别规则，进而利用不同海岸类型所包含的地物的差异，提出了基岩、人工建筑、人工养殖、沙滩、淤泥五种海岸类型的自动识别规则。为了验证本文提出的海岸类型自动识别规则的有效性，选取了 50 景 80×80 大小的海岸区域进行识别验证，实验结果表明，对于基岩海岸和人工海岸能够高精度的识别出来，砂质海岸和淤泥质海岸的识别精度略低，但平均识别精度达到 86%。这说明对于地物特征明显的海岸类型，本文提出的海岸类型自动识别方法能够代替人工目视识别，具有一定的应用价值。

本文中地物识别规则是以 2013-04-21 获取的 Landsat 8 影像为例分析得出的，对于其他时间获取的影像，规则应用的精度或许不高。因此，建议在进行海岸类型识别之前，应在所采用的影像数据上重新选取海水、陆地、植被、养殖区、淤泥和沙滩样本，采用 C4.5 决策树算法重新进行分析，以便得到适合待检测影像的地物分类规则，然后再采用本文提出的海岸类型识别规则进行海岸类型识别。

参考文献：

[1]　WANG C Y, ZHANG J, SONG P J. An intelligent coastline interpretation of several types of seacoasts from TM/ETM plus images based on rules[J]. Acta Oceanologica Sinica, 2014, 33(7): 89-96.

[2]　庄翠蓉. 厦门海岸线遥感动态监测研究[J]. 海洋地质动态, 2009, 25(4): 13-17.

[3] 马小峰，赵冬至，邢小罡，等. 海岸线卫星遥感提取方法研究[J]. 海洋环境科学，2007,26(2)：185-189.

[4] WANG C Y, ZHANG J, MA Y. Coastline interpretation from multispectral remote sensing images using an association rule algorithm[J]. International Journal of Remote Sensing, 2010, 31(24): 6409-6423.

[5] ZHANG T, YANG X M, HU S S. Extraction of coastline in aquaculture coast from multispectral remote sensing images: object-based region growing integrating edge detection[J]. Remote Sensing, 2013,5(9): 4470-4487.

[6] 齐宇，任航科. 基于厦门岛的海岸线自动提取方法研究[J]. 城市勘测，2012，(5)：75-78.

[7] WANG C Y, ZHANG J, MA Y. Coastal land covers classification of igh-resolution images based on dempster-shafer evidence theory[C]// International Conference on Computer Science and software engineering. Wuhan:IEEE,2008:1061-1064.

[8] WANG L, SHAO F J, SUN R C, et al. The automatic classification of seacoast types from multispectral image based on data mining[C]// International Conference on Computer and Information Science, Safety Engineering. Shanghai:IEEE Conference Publishing Sevice, 2012.

[9] 徐涵秋，唐菲. 新一代 Landsat 系列卫星:Landsat 8 遥感影像新增特征及其生态环境意义[J]. 生态学报，2013，33(11)：3249-3257.

[10] 邵峰晶，于忠清，王金龙，等. 数据挖掘原理与算法[M]. 北京：科学出版社，2009.

Automatic Recognition of Coast Types in Landsat 8 OLI Images by Using Decision Tree Algorithm

WANG Chang-ying[1], TAN Meng[2]

(1. *College of Information and Engineer, Qingdao University*, Qingdao 266071, China;
2. *North China Sea Data & Information Service of the State Oceanic Administration, SOA*, Qingdao 266061, China)

Abstract: For recognizing the following four types of coasts including bedrock coast, artificial coast, sandy coast and mud coast, the Landsat 8 OLI multispectral images are analyzed by means of C4.5 decision tree algorithm of data mining technique. Firstly, the rules for recognizing the land features such as seawater, land, vegetation, aquaculture, mud and beach which are contained in the four coast types are worked out, and then the rules for automatically recognizing five types of coasts including bedrock coast, man-made structure coast, artificial aquaculture coast, sandy coast and mud coast are respectively proposed according to the differences in land features contained in different types of coasts. The experimental results show that the accuracy of the proposed automatic recognition rules is 100% for the artificial and the bedrock coasts with obvious land features and greater than 80% for the mud and the artificial aquaculture coasts with similar spectra.

Key words: Landsat 8; coast type; automatic recognition; data mining; C4.5 decision tree