우연수림분류기에 의한 위성화상분류방법

전금성, 리금수

론문에서는 통계부문에서 쓰이는 우연수림분류기를 리용하여 위성화상을 분류하는 방법에 대하여 소개하고 이 분류결과의 정확도를 지지벡토르분류기에 의한 분류결과와 대비하여 고찰하였다.

1. 우연수림분류기의 원리

우연수림(RF: Random Forests)분류기에 의한 분류리론은 2001년 처음으로 통계분야에서 발표되였다. 2012년 우연수림분류기에 대한 코드가 Fortran언어로 공개되였으며 2013년에는 R언어로 공개되였다.[1] 여기에서 변수의 중요성과 오차 등의 지표가 계산되였다.

우연수림분류기는 많은 결정나무분류모임($h(X,\ \boldsymbol{\theta}_k), k=1, \cdots,\ n$)으로 이루어진 분류모형이다. 여기서 $\boldsymbol{\theta}_k$는 서로 독립인 벡토르량이다. 변수 $\boldsymbol{X}$가 주어진 상태에서 이 분류모형은 결과들의 반복회수에 따라 최종분류결과를 얻어낸다.[2] 그 기본원리는 다음과 같다.

우선 육안으로 대상들을 선택한 모임에서 bootstrap표본선택법에 의하여 n개의 표본을 뽑아 표본모임을 만든다. 이때 뽑은 표본개수는 선택한 매 대상의 화소수의 3분의 2정도 되여야 한다. 매 표본모임에 대하여 결정나무를 만든다. 그러면 k개의 결정나무가 얻어지며 이 나무들에 의하여 k개의 결과가 얻어지게 된다.

k개의 결과중에서 제일 많은 분류결과가 최종결과로 되게 된다.

$$H(\boldsymbol{X}) = \arg(\max_y \sum_{i=1}^{k} I(h_i(\boldsymbol{X}) = Y)) \tag{1}$$

우연수림을 이루는 결정나무를 만드는 알고리듬에는 ID3, C4.5, CART, PUBLIC 등이 있다.[3]

론문에서는 개별적인 결정나무를 만드는 알고리듬으로 CART(Classification And Regression Tree)를 리용하였다.

CART알고리듬은 *Gini*지표에 의하여 아지가 갈라지는 2진결정나무알고리듬이다. 마디점에서 가지가 갈라질 때 *Gini*지표값을 계산하면 그에 따라 좌, 우로 가지가 갈라지게 된다.

① *Gini*지표값계산

$$Gini(S) = 1 - \sum_{i=1}^{m} P_i^2 \tag{2}$$

여기서 P_i는 무리 s_i가 모임 S에서 나타날수 있는 확률이다.

② 마디점에서 갈라질 때 리용하는 $Gini_{split}$지표값계산

$$Gini_{split}(S) = \frac{|s_1|}{|S|} Gini(S_1) + \frac{|s_2|}{|S|} Gini(s_2) \tag{3}$$

만일 모임 S가 자식모임 s_1과 s_2로 갈라진다면 이때 계산되는 많은 $Gini_{split}(S)$ 값들가운데서 최소값을 취하여 모임의 분할을 진행한다.

이러한 원리에 의하여 결정나무를 얻는다.

2. 우연수림분류기에 의한 위성화상의 분류

론문에서는 널리 알려진 Landsat TM화상을 가지고 화상분류를 진행하였다. 화상을 논, 밭, 과수밭, 산림, 수역, 라지로 분류하였다.

우연수림분류기분류에서 리용하는 파라메터에는 나무의 개수, 자식나무특징개수, 나무높이, 최소표본개수, 정확도, 턱값이 있다.

나무의 개수가 클수록 분류정확도는 높아지지만 계산량이 많아지게 된다.

이 실험에서는 나무개수를 50개로 하였다. 자식나무특징개수는 일반적으로 총특징개수의 2차뿌리를 취하는데 론문에서 리용한 자료가 6개 파장대역을 가지기때문에 6의 2차뿌리의 근사값 2를 택하였다.

나무높이가 높을수록 계산량이 많아지게 되는데 론문에서는 경험적으로 10을 정하였다. 나무잎의 최소표본개수는 리용한 표본개수가 적은것으로 하여 3개로 하였다. 분류에 리용한 화상자료는 그림 1과 같다.

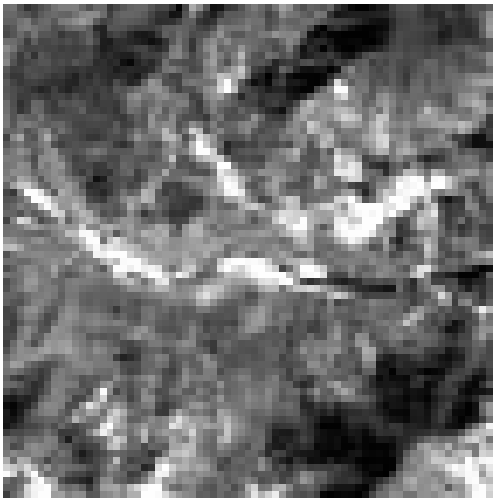

그림 1. 분류에 리용한 화상

이 화상의 분해능은 30m로서 공간분해능이 비교적 낮기때문에 표본자료를 육안으로 정확히 선택하는것은 불가능하다. 따라서 이미 가지고있는 벡토르화상을 가지고 이것을 참고로 표본자료를 선택하였다. 화상의 크기는 71×72이다.

론문에서는 우연수림분류기에 의한 화상분류와 지지벡토르분류기에 의한 화상분류를 각각 진행하였다. 이때 분류에 리용한 표본자료는 표 1과 같다. 표본자료는 벡토르화상을 라스터자료로 넘겨 얻을수 있다. 그렇게 하면 벡토르자료가 라스터자료로 넘어가면서 경계선에서 불정확성이 산생되게 된다.

표 1. 분류에 리용한 표본자료

분류이름	표본화소개수/개	검사화소개수/개
논	72	41
밭	54	29
과수밭	39	26
산림	67	55
수역	68	41
라지	49	37

실례로 논과 밭의 경계선이 벡토르자료로 넘어가면 선을 이루는 매점은 각각 1개의 화소점으로 넘어가게 되는데 1개 화소는 현지의 30m×30m지역을 의미한다.

이것은 1개 밭의 크기와 거의나 같다. 즉 벡토르－라스터변환의 정확도문제로 하여 이러한 표본자료선택은 불합리하다.

따라서 수동적인 방법으로 벡토르자료와 위성화상자료를 겹친 상태에서 벡토르자료에 기초하여 표본자료를 선택하였다. 같은 방법으로 검사자료도 선택하였다.

실험에서 TM화상의 1, 2, 3, 4, 5, 7대역을 리용하였다.

분류결과에 대한 분석에서는 ENVI프로그람의 Kappa결수분석을 리용하였다.

분류결과는 그림 2와 3과 같다.

그림 2. 지지벡토르분류기에 의한 화상분류결과

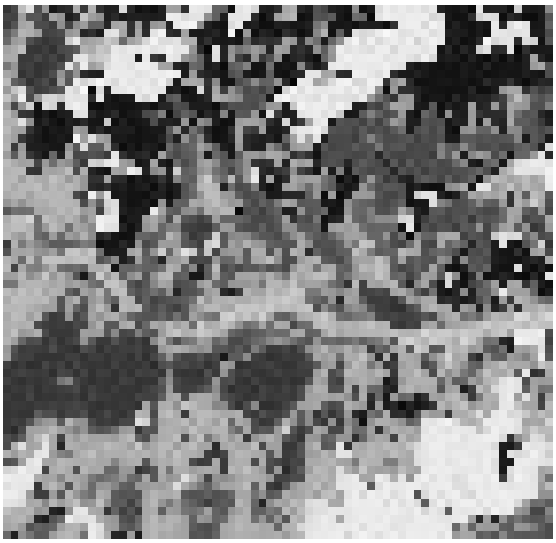

그림 3. 우연수림분류기에 의한 화상분류결과

Kappa지표에 의한 분류결과는 표 2와 같다.

표 2. Kappa지표에 의한 분류결과

지표	우연수림분류기	지지벡토르분류기
Kappa	0.829 8	0.803 2
총정확도	86.026 2%	83.842 8%

검사에서 보면 우연수림분류기에 의한 분류에서 Kappa지표가 지지벡토르분류기보다 0.026% 높았으며 총 정확도는 2.18% 높았다.

결과적으로 화상에 대한 분류에서 우연수림분류기에 의한 분류정확도가 지지벡토르분류기에 의한 검사결과에 비하여 우월하다는것을 알수 있다.

맺 는 말

통계부문에서 쓰이는 우연수림분류기를 리용하여 위성화상분류를 진행하면 지지벡토르분류기에 의한 분류보다 더 정확하다는것을 검증하였다.

참 고 문 헌

[1] D. Richard Cutler et al.; Ecological Society of America, 88, 11, 2783, 2007.

[2] 李欣海; 应用昆虫学报, 50, 4, 1190, 2013.

주체108(2019)년 1월 5일 원고접수

The Classification of Satellite Images Based on Random Forest Classifier

Jon Kum Song, Ri Kum Su

In this paper we studied the classification of satellite images based on Random Forest Classifier, which was used in statistics, conducted an analysis of the precision via confusion matrix, and compared the precision with the classification by SVM.

Key words: Random Forest Classifier, decision tree