ISODATA와 퍼지 C-Means를 이용한 감독 분류의 성능 향상에 관한 연구

전영준° 김진일
동의대학교 컴퓨터공학과
j4017@chol.com, jikim@dongeui.ac.kr

A Study on Improving Performance of Supervised Classifier using ISODATA and Fuzzy C-Means Clustering Method

Youngjoon Jeon° Jinil Kim
Dept. of Computer Engineering, Dongeui University

요 약

본 논문에서는 위성영상의 감독 분류에 대한 성능 개선을 위하여 ISODATA와 퍼지 C-Means 클러스터링 기법을 이용한 베이시안 최대우도 분류방법을 제안하였다. 본 연구에서는 ISODATA 클러스터링 기법을 이용하여 각각의 분류항목별로 분광특징에 따라 분석가가 선정한 훈련 데이터를 분할하여 새로운 훈련 데이터를 선정함으로써 분류항목별 훈련데이터의 분광적인 특징에 관계없이 분류를 수행할 수 있도록 하였다. 그리고 새롭게 선정된 훈련 데이터를 이용하여 퍼지 C-Means 클러스터링을 수행하고 그 결과를 베이시안 최대우도 분류기법의 사전확률로 이용함으로써 위성영상의 감독 분류에 대한 성능을 개선할 수 있는 방법을 제안한다. 제안된 기법은 Landset TM 위성영상을 이용하여 그 적용성을 시험하였다.

1. 서 론

원격탐사에 의한 위성영상 정보는 토지의 체계적인 활용, 환경오염의 감시 및 통제, 지도제작 등에 요구되는 복잡 다양한 정보를 신속하고 정확하게 분류, 해석하는데 있어서 커다란 잠재력을 가지고 있는 최첨단 과학기술로 각광 받고 있으며, 그 활용 분야는 환경에서부터 토목, 농업, 해양, 지질, 임업, 수산업 등 각 분야에 걸쳐 급속도로 확산되고 있다. 원격탐사 위성영상의 분류방법에는 감독 분류와 무감독 분류가 있다[1]. 감독 분류 기법은 표본집단과 같은 사전정보 없이 영상을 구성하는 화소값의 공간적, 분광적 특성만을 이용하여 분류 작업을 수행하는 기법이다. 이 방법은 순수한 통계 처리의 기법에 의해 수행된다는 점에서 군집화(clustering)라고 불리며, 분류할 군집의 개수와 각 분광 군집 사이의 한계 거리만을 지정해 줌으로써 작업이 가능하다. 무감독 분류 기법에는 순차군집분류 기법, K-Means 군집분류기법, ISODATA 기법, 퍼지 C-Means 군집분류기법 등이 있다. 감독 분류는 분석가가 영상 내에서 알고 있는 픽셀들의 영역을 지정하면, 각 군집의 중심, 밴드간 공분산 등을 이용하여 자료내의 모든 픽셀을 분석하여 가장 유사한 분포특성을 가지는 군집에 할당시키는 과정이다. 감독 분류 기법에는 평행육면체 기법, 최소거리 기법, 최대우도 분류기법[2] 등이 있다. 위성영상 분류의 최근 연구 동향은 1980년대 후반에 들면서 인공지능, 퍼지, 신경망 이론이 본격적으로 등장하면서 기존의 통계적 이론에서 보다 개선된 제안들이 속속 소개되고 있다. 이에는 퍼지기법을 적용하여 분류의 개선을 시도한 연구가 있으며, 신경망 이론의 위성이미지 분류에의 적용에 대한 연구가 다양하게 이루어지고 있다.[3][4] 본 논문에서는 위성영상의 분류의 정확성을 개선시키기 위하여 ISODATA를 이용하여 분류 항목별 훈련 데이터를 분광특징에 따라 세부항목으로 분리하고, 이를 이용한 퍼지 C-Means의 분류 결과를 베이시안 최대우도 분류기의 사전확률로 이용하여 분류를 수행하여 감독 분류 성능을 향상시키는 방법을 제안하였다.

2. 퍼지 베이시안 최대우도 분류 시스템

2.1 ISODATA 분류기법을 이용한 분류항목의 선정

위성영상의 감독 분류는 특정 지역을 분류하기 위한 사전 훈련 데이터가 필요하다. 훈련 과정의 전반적인 목표는 특정 영상에서 구분되는 모든 토지피복 종류에 대한 분광반응패턴을 설명할 수 있는 통계집단을 수집하는 것이다. 좋은 분류결과를 산출하기 위해, 훈련 데이터는 대표성과 완벽성을 가져야 한다. 본 연구에서는 입력영상에 대하여 분석가가 먼저 분류항목별로 훈련 데이터를 여러 영역을 선정한다. 그리고 분류 항목별로 선정된 훈련 데이터에 대하여 각각 ISODATA 분류를 수행하여 새로운 훈련 데이터를 선정한다. 이 과정은 분류항목별로 선정된 훈련 데이터를 분광특징에 따라 세부분류항목으로 세분화하여 새로운 훈련 데이터로 선정하는 것이다. 이렇게 함으로서 훈련 데이터는 작은 값의 밴드별 분산값을 가지며, 정규분포 형태의 자료 분포를 보이게 된다. ISODATA 분류 알고리즘은 매 반복 단계마다 표본의 평

균을 군집의 중심으로 정하며, 군집의 삭제, 분리, 병합을 통해 자기 조직화가 가능하며, 고정된 수의 군집들을 처리하는 것이 아니라, 최종 군집의 개수는 사용자의 요구 군집 개수에 의존하지 않고 적당한 군집의 개수를 생성한다는 특징을 가지고 있다. ISODATA 알고리즘은 각 군집마다 허용되는 샘플들의 최소 개수, 병합이 일어나지 않는 군집중심 사이에 허용되는 최소거리, 군집의 분리를 조절하는 파라미터, 각 반복에서의 군집 병합의 최대 수, 알고리즘의 반복의 최대 수를 입력 값으로 한다. 그리고 본 연구에서는 ISODATA 군집화를 위해 픽셀과 군집 중심과의 거리는 마하라노비스 거리를 사용하였다. 마하라노비스 거리는 분산의 차이에 대해 각 축간 모집단의 분포 상관을 고려한 보정을 행하고 있는 거리이다.

2.2 퍼지 베이시안 분류기법

분석가에 의하여 분류항목별 훈련 데이터를 선정한 후 이를 ISODATA 분류기법을 이용하여 각각의 분류항목별로 분광특징에 따라 훈련 데이터를 세분화하여 새로운 훈련 데이터를 선정하였다. 분류항목별로 세분화된 훈련 데이터로부터 평균값을 구하여 퍼지 C-Means의 중심값으로 설정하고 각 중심값에 대한 소속도를 구하여 분류를 수행하였다. 퍼지 C-Means의 분류결과를 베이시안 최대우도 분류기의 사전확률로 이용하여 분류를 수행하였다.

퍼지 C-Means 분류기법은 목적함수기반(Objective Function Based) 퍼지 규칙을 사용한다.[5] 퍼지 C-Means분류기법은 각 화소들이 군집에 소속될 소속도와 군집중심을 생성한다.

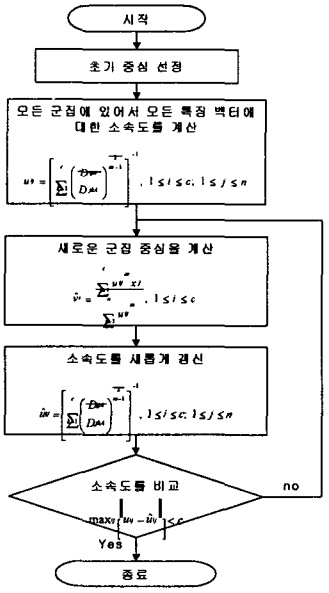

그림 1. 퍼지 C-Means 분류과정

퍼지 C-Means 군집분류기법에서 사용하는 목적함수는 J는 다음과 같다.

$$J_m(U,V;X)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}{}^{m}\left\|x_j-v_i\right\|^2$$

위의 목적함수 J의 값을 최소화시키는 u_{ij}와 v_i를 구하는 것이 목표이다. 여기에서 m은 퍼지 식별자라고 하고 만약 m=1인 경우는 HCM(Hard Fuzzy C-Means)이며, 퍼지 C-Means 군집분류 기법에서는 $1<m<\infty$의 값을 사용하며 일반적으로 m=2값을 사용한다.

$$M_{fcn}=\left\{U\in R^{nc}\middle|u_{ij}\in[0,1]\forall i,j;\ 0<\sum_{j=1}^{n}u_{ij}<n\ \forall i,\ and\ \sum_{i=1}^{c}u_{ij}=1\ \forall j\right\}$$

본 논문에서는 퍼지 C-Means의 반복수행을 하지 않고 ISODATA에 의하여 새롭게 선정된 훈련 데이터의 중심값을 이용하여 한번만 수행한다.

퍼지 베이시안 최대우도 분류기법은 퍼지 C-Means 분류 결과로부터 예측되는 분류항목의 분포로부터 사전확률을 추출하여 각 분류항목별 가중치를 지정하여 베이시안 최대우도 분류를 수행하게 된다. 새롭게 선정된 훈련 데이터를 이용하여 퍼지 베이시안 최대우도 분류를 수행한다. 일반적으로 사전확률을 가지는 베이시안 최대우도 분류기법은 판별식의 $D_i(x)$값의 값을 최대화되는 분류항목으로 픽셀을 분류하는 과정이다. 이때 $P(w_i)$는 사전확률을 의미하는 것으로서 본 연구에서는 퍼지 C-Means 분류결과를 이용하였다.

$$D_i(X)=\ln P(w_i)-\frac{N}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|-\frac{1}{2}(X-U_i)^T\Sigma_i^{-1}(X-U_i)$$

위 식은 다음과 같이 나타낼 수 있다.

$$D_i(X)=\ln P(w_i)-\frac{N}{2}\ln 2\pi-\frac{1}{2}\begin{vmatrix}c_{i11}&c_{i12}&\cdots&c_{i1N}\\c_{i21}&c_{i22}&\cdots&c_{i2N}\\\vdots&\vdots&\cdots&\vdots\\c_{iN1}&c_{iN2}&\cdots&c_{iNN}\end{vmatrix}-A$$

$$A=\frac{1}{2}\left[\begin{pmatrix}x_{1j}\\x_{2j}\\\vdots\\x_{Nj}\end{pmatrix}-\begin{pmatrix}\mu_{1i}\\\mu_{2i}\\\vdots\\\mu_{Ni}\end{pmatrix}\right]^T*\begin{vmatrix}c_{i11}&c_{i12}&\cdots&c_{i1N}\\c_{i21}&c_{i22}&\cdots&c_{i2N}\\\vdots&\vdots&\cdots&\vdots\\c_{iN1}&c_{iN2}&\cdots&c_{iNN}\end{vmatrix}^{-1}*\left[\begin{pmatrix}x_{1j}\\x_{2j}\\\vdots\\x_{Nj}\end{pmatrix}-\begin{pmatrix}\mu_{1i}\\\mu_{2i}\\\vdots\\\mu_{Ni}\end{pmatrix}\right]$$

N :밴드 수

X : 데이터 벡터

U_i : 분류항목i에 대한 N개 밴드의 평균벡터

$\_$: 분류항목 i의 공분산 행렬

C_{kl} : 두 개의 밴드들 사이의 공분산(k와 l밴드)

$$C_{kl}=\frac{\left(\sum_{j}^{n}(x_{kj}-u_k)(x_{lj}-u_l)\right)}{n-1}$$

밴드 수 : k=1, 2, ……, N

밴드 수 : l=1, 2, ……, N

픽셀 값 : i=1, 2, ……, N

$|\_|$: 공분산 행렬 $\_$ 의 determinant

$\_^{-1}$: $\_$ 의 역행렬

$(X-U_i)^T$: 벡터 $(X-U_i)$의 전치행렬

각 훈련 데이터에 의해 분류가 완료되면 세부분류항목에 따른 분류결과를 원래의 분석가가 설정한 분류항목으로 재설정한다.

3. 실험 및 고찰

본 논문에서는 30m×30m의 공간해상도를 가지는 Landsat TM 위성영상을 입력으로 제안한 분류 알고리즘의 정확도를 실험하였다. 제안된 알고리즘의 정확성을 테스트하기 위해서 사용된 이미지는 미국의 Purdue 대학에서 수집한 데이터로서 1986년에 관측된 미국의 *Indiana*의 *Tippecanoe County* 지역에 대한 Landsat TM 영상이다.[6] 영상의 크기는 가로와 세로의 크기가 169×169 화소이고, 밴드 수는 7개이며, 파일형식은 BIL 포맷이다. 하나의 화소는 8bit로 구성되어 있다. 분석가에 의하여 corn, soybean, wheat, alfalfa/oats, pasture의 5개의 분류항목을 정의하고, Purdue 대학에서 수집한 실측데이터에 의하여 각 분류항목별로 훈련 데이터를 선정하였다. 선정된 훈련 데이터의 각 군집에 대한 중심을 ISODATA 군집분류기법의 초기군집 중심으로 이용하여 분광특징별로 다수의 군집으로 분리하여 새롭게 훈련 데이터를 생성하였다. 새롭게 생성된 훈련 데이터는 전체 분류항목에 대하여 8개의 군집을 형성하였으며, ISODATA 분류를 수행하여 생성된 군집은 비교적 적은 값의 밴드별 분산값을 가지며, 정규분포 형태의 자료분포를 보여 주었다.

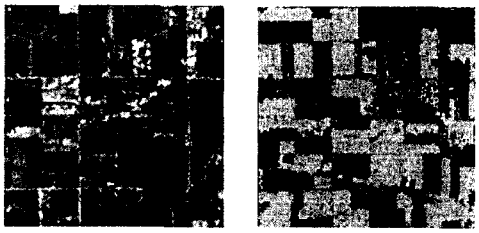

그림 2. 원래 영상(좌)과 퍼지 베이시안 분류결과 영상(우)

표 1. 퍼지 베이시안 분류에 대한 confusion matrix

분류항목	Fuzzy Bayesian maximum likelihood					Total
	corn	soybean	wheat	alfalfa/oats	pasture	
corn	7211	696	21	251	14	8,193
soybean	854	6465	34	499	4	7,856
wheat	23	129	163	35	1	351
alfalfa/oats	285	386	31	966	1	1,669
pasture	2	3	2	5	22	34
total	8,375	7,679	251	1,756	42	18,103

일반적인 방법의 최대우도 분류를 수행한 결과는 전체 분류 정확도의 퍼센티지가 70% 이었으며, 본 연구에서의 퍼지 베이시안 최대우도 분류는 표1에 의하여 전체 정확도가 82%로 성능이 향상되었음을 보여주었다. 특히 분산값이 크게 나타나는 지역의 분류에 좋은 효과를 나타내었다.

4. 결론

본 논문에서는 베이시안 최대우도 분류기법을 이용한 위성영상의 감독 분류 정확도 향상을 위하여 ISODATA 분류기를 이용한 훈련 데이터의 선정 및 퍼지 C-Means를 사전확률로 하는 베이시안 최대우도 분류기법에 대해서 연구하였다. 본 연구에서는 위성영상 이미지의 분류에 있어, 분석가에 의하여 분류항목별 훈련 데이터를 선정한 후 훈련데이터에 대하여 ISODATA 분류를 수행하여, 분류항목별 훈련 데이터를 새롭게 설정한다. 새롭게 설정된 각각의 훈련 데이터로부터 평균값을 구하여 퍼지 C-Means의 중심값으로 설정하고 각 분류항목에 대한 소속도를 구하여 분류를 수행한다. 퍼지 C-Means의 분류 결과를 베이시안 최대우도 분류기의 사전 확률로 이용하여 분류를 수행하였다. 최종적으로 분석가가 선정한 분류항목에 따라 분류결과를 재설정함으로써 분석가가 의도하는 항목에 대한 분류결과를 얻을 수 있다. Landsat TM 영상을 이용한 실험으로 일반적인 베이시안 최대우도 분류기법보다 분류의 성능을 개선할 수 있었으며, 또한 분석가가 선정한 훈련데이터의 분광적인 특징에 관계없이 분류를 수행할 수 있었다.

향후 연구과제는 다양한 종류의 위성영상들에 대하여 제안한 분류방법의 적용성에 대한 연구가 뒤따라야 하겠다.

참고문헌

[1] John A. Richards, Remote Sensing Digital Image Analysis : An Introduction, Second, Revised and Enlarged Edition, pp. 229-262, Springer-Verlag, 1994.

[2] B.Gorte and A. Stein. Bayesian classification and class area estimation of satellite images using stratification. IEEE Trans. On Geoscience and Remote Sensing, 36(3):303, 1998

[3] Kent, J.T. and Mardia, K.V., "Spatial classification using fuzzy membership models", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.10, No.5, pp.659-671, 1988.

[4] Heermann,P.D. and Khazenie,N., "Classification of multispectral remote sensing data using a back- propagation neural network" , IEEE Trans. on Geosci. and Remote Sensing, Vol.30, NO.1, pp.81-88, Jan., 1992.

[5] Frank Hoppner, Frank Klawonn, Rudolf Kruse, Thomas Runkler FUZZY CLUSTER ANALYSIS Methos for Classification, Data Analysis and Image Recognition, John Wiley & Sons Ltd, pp. 1-59, 1999.

[6] Laboratory for Applications of Remote Sensing, Purdue University : http://www.lars.purdue.edu.