이 성 건 교수지도

석사학위 청구논문

위치모수와 척도모수의 동시검정법을 이용한 이미지 픽셀의 분류

2016

성신여자대학교 대학원

통계학과

이 민 주

위치모수와 척도모수의 동시검정법을 이용한 이미지 픽셀의 분류

이 성 건 교수지도

이 논문을 석사학위논문으로 제출함

2015년 11월

성신여자대학교 대학원

통계학과

이 민 주

인 준 서

이민주의 석사학위 논문으로 인준함.

2015년 11월

심사위원장 ..... (인)

심 사 위 원 ..... (인)

심 사 위 원 ..... (인)

성신여자대학교 대학원

논 문 개 요

본 논문에서는 새로운 이미지(image) 픽셀(pixel)값의 정분류율을 높일 수 있는 분류방법에 대해 제안하고자 한다. 이미지 픽셀의 분류는 의료분야에서 활발히 사용 중이다. 예를 들면, 정상인과 환자의 의료용 이미지(x-ray, CT, MRI 등)를 대조하여 의학적 진단을 내릴 수 있으며, 해부구조의 연구, 얼굴인식분야에서도 사용한다. 그 밖에 천문학, 농림업, 컴퓨터 비전(computer vision) 분야에서도 사용된다.

픽셀이란 컴퓨터에서 주소화될 수 있는 화면의 가장 작은 단위로, 작은 점의 행과 열로 이루어져 있는 화면의 작은 점 각각을 이르는 말이다. 이미지는 픽셀의 유한한 집합으로, 서로 다른 영역을 대표하는 픽셀값의 그룹으로 구현된다. 이미지에서 특정 부분이나 색을 나타내는 픽셀들은 서로 비슷한 값을 가진다. 이러한 성질을 이용하여 유사한 값을 가지는 픽셀들을 동일한 집단으로 여겨 새로운 픽셀의 집단을 할당할 수 있다.

최근 Liao & Akritas(2007)은 가설검정 기반 이미지 픽셀 분류방법을 새롭게 제안하였으나, 그들이 제안한 방법은 p-값이 너무 작을 경우 오분류율이 높다는 단점이 있다. 이 단점을 보완하기 위해 Ghimire & Wang(2012)은 p-값이 아주 작은 경우에 평균과의 거리를 이용하여 픽셀값을 분류하는 방법을 고안하였다. Liao & Akritas의 방법(이하 LA 방법)보다 오분류율이 줄었지만, 위치모수만을 고려한 Ghimire & Wang의 분류방법(이하 GW 방법) 역시 집단별 분산에 차이가 있을 때 오분류율이 높다는 단점이 있다.

따라서 본 논문에서는 집단별 분산이 다를 경우, 위치모수(location

parameter)와 척도모수(scale parameter)의 동시검정법을 이용한 가설검정기반 이미지픽셀 분류방법을 새롭게 제안하고자 한다. 이미지 픽셀값의 위치모수와 척도모수를 동시에 고려하였기 때문에 정분류율이 선행연구의 방법보다 높을 것으로 예상하였다. 모의실험을 통하여 가설검정을 기반으로 한 LA 방법, 위치모수만을 고려한 GW 방법, 위치모수와 척도모수의 동시검정법인 Kolmogorov- Smirnov 검정(Smirnov, 1939), Cramér-von Mises 검정(Anderson, 1962), Lepage 검정(Lepage, 1971), Cucconi 검정(Cucconi, 1968) 중 어떤 방법의 정분류율이 높은지 확인하였다.

모의실험 결과 두 집단의 분산이 동일할 때, GW 방법이 우수했으나 위치모수와 척도모수의 동시검정법 역시 이미지 픽셀의 정분류율이 높음을 확인하였다. 두 집단의 분산이 다를 때는 위치모수와 척도모수의 동시검정법을 이용한 분류의 정분류율이 더 높음을 확인하였다. 특히 Cucconi 검정을 이용한 이미지픽셀의 분류가 가장 높은 정분류율을 보였다.

의료분야, 얼굴인식 분야 등 실제 이미지 픽셀 분류에 사용되는 이미지와 분류에 관심 있는 집단은 평균과 분산이 다른 경우가 많기 때문에 위치모수만 고려한 GW 방법을 이용하는 것보다는 본 연구에서처럼 위치모수와 척도모수의 동시검정법을 이용하여 이미지 픽셀을 분류하는 것이 더 정분류율을 높일 수 있을 것이다.

목 차

표 목 차

그 림 목 차

제 1 장 서론

본 논문에서는 새로운 이미지(image) 픽셀(pixel)값의 정분류율을 높일 수 있는 분류방법에 대해 제안하고자 한다. 이미지 픽셀의 분류는 의료분야에서 활발히 사용 중이다. 예를 들면, 정상인과 환자의 의료용 이미지(x-ray, CT, MRI 등)를 대조하여 의학적 진단을 내릴 수 있으며, 해부구조의 연구, 얼굴인식분야에서도 사용한다. 그 밖에 천문학, 농림업, 컴퓨터 비전(computer vision) 분야에서도 사용된다.

픽셀이란, 컴퓨터에서 주소화될 수 있는 화면의 가장 작은 단위로, 작은 점의 행과 열로 이루어져 있는 화면의 작은 점 각각을 이르는 말이다. 이미지는 픽셀의 유한한 집합으로, 서로 다른 영역을 대표하는 픽셀값의 그룹으로 구현된다. 이미지에서 특정 부분이나 색을 나타내는 픽셀들은 서로 비슷한 값을 가진다. 이러한 성질을 이용하여 유사한 값을 가지는 픽셀들을 동일한 집단으로 여겨 새로운 픽셀의 집단을 할당할 수 있다.

흔히 이미지 픽셀 분류에 사용되는 통계적 방법은 선형 판별분석(LDA, linear discriminant analysis, Hastie et al., 2009), 이차 판별분석(QDA, quadratic discriminant analysis, Hastie et al., 2009), 의사결정나무(decision tree, Breiman et al., 1984) 등이다. 컴퓨터를 기반으로 한 분류방법은 k-근접이웃(k-nearest-neighbor)방법, 신경망 분석(neural network), 서포트 백터 머신(SVM, support vector machine) 등이 있다(Vapnik, 1982).

최근 Liao & Akritas(2007)은 가설검정 기반 이미지 픽셀 분류방법을 새롭게 제안하였으나, 이 방법(이하 LA 방법)은 p-값이 너무 작을 경우 오분류율이 높다는 단점이 있다. 이 단점을 보완하기 위해

Ghimire & Wang(2012)은 p-값이 아주 작은 임계값(ϵ)보다 큰 경우(p-값>ϵ)에는 LA 방법을 사용하지만, p-값이 아주 작은 임계값보다 작은 경우(p-값<ϵ) 평균과의 거리를 측정하여 픽셀값을 분류하는 방법(이하 GW 방법)을 고안하였다. GW 방법은 LA 방법보다 오분류율이 줄었지만, 위치모수만을 고려한 GW 방법도 집단별 분산에 차이가 있을 때 오분류율이 높다는 단점이 있다. 분산에 대한 고려가 없이 단순히 평균과의 거리로 분류하기 때문에, 한 집단의 분산이 매우 커서 다른 픽셀값과의 접점이 생기는 경우, 오분류율이 높게 나타났다.

따라서 본 논문에서는 위치모수와 척도모수의 동시검정법을 이용한 가설검정기반 이미지픽셀 분류방법을 새롭게 제안하고자 한다. 이미지 픽셀값의 위치모수와 척도모수를 동시에 고려하였기 때문에 정분류율이 선행연구의 방법보다 높을 것으로 예상하였다.

모의실험을 통하여 가설검정을 기반으로 한 LA 방법, 위치모수만을 고려한 GW 방법, 위치모수와 척도모수의 동시검정법(location-scale test)인 Kolmogorov - Smirnov 검정(Smirnov, 1939, 이하 K-S 검정), Cramér - von Mises 검정(Anderson, 1962, 이하 CVM 검정), Lepage 검정(Lepage, 1971), Cucconi 검정(Cucconi, 1968)을 이용하여 이미지 픽셀을 분류하였다. 그리고 총 여섯 가지 분류방법의 정분류율을 비교하여 어떤 검정이 더 좋은 방법인지 살펴보았다.

본 논문의 순서는 다음과 같다. 제 2장에서는 이미지 픽셀과 분류문제, 개념들에 대해 설명하고, 제 3장에서는 새로운 픽셀값을 분류하기 위하여 본 논문에서 사용한 여섯 가지 방법을 소개한다. 제 4장에서는 여섯 가지 방법을 이용하여 모의실험을 설계하고 수행한다. 마지막으로 제 5장에서는 연구의 결론과 향후 연구방향을 제시하고자 한다.

제 2 장 이미지 픽셀

이 절에서는 논문을 이해하기 위하여 필요한 몇 가지 개념들에 대해 소개하고자 한다. 우선, 픽셀값으로 이루어진 이미지를 0과 1사이의 값으로 수치화하고 3차원으로 구성된 이미지의 차원을 축소하는 방식을 살펴본다. 또한, 이미지를 간소화하기 위하여 이미지를 대표하는 집단(class)과 훈련 자료(training data), 검증자료(test point)에 대해 알아본다.

2.1. 이미지 픽셀값의 수치화

본 절에서는 이미지 픽셀값을 수치화 하고 3차원의 RGB값을 축소하여 1차원인 흑백이미지 픽셀값으로 변환하는 방법에 대해 설명하고자 한다.

컴퓨터는 색을 숫자로 인식하고 숫자들의 집단을 통하여 이미지를 출력한다. 색 공간(color space)을 정의하는 모델은 매우 다양하다. 일반적으로 컴퓨터에서 널리 사용되고 있는 색공간은 RGB 모델(red-green-blue color model)이며, 그 밖에 잉크색을 기준으로 하여 인쇄시스템에서 사용되는 CMYK 모델(cyan-magenta-yellow- black), 색상의 정보를 밝기와 색으로 구분하여 영상압축 분야에서 사용되는 YUV 모델 등이 있다. 본 논문에서는 RGB 모델과 YUV 모델간 변환 방법을 이용하여 데이터를 얻고자 한다. 먼저 두 모델은 어떤 특성을 가지고 있는지 각각 살펴보자.

RGB 모델은 적색, 녹색, 청색을 혼합하여 원하는 색을 만드는 가색 방식을 사용한다. 적색과 녹색, 청색 채널로 이미지픽셀은 다양한 색을 출력할 수 있다. 세 개의 채널은 0과 255사이의 값을 가질 수 있다. 세

가지 차원으로 색상을 표현하기 때문에 세밀하고 다양하게 색 표현이 가능하지만 용량이 커진다는 단점이 있다.

YUV 모델은 이미지의 명암을 나타내는 휘도신호(luminance signal) Y와 적색, 녹색, 청색의 식 신호의 차이를 나타내는 색차신호(color difference signal) U, V로 색상을 표현한다. 휘도신호 Y는 16부터 235 사이의 값을 가지며, 색채신호 U, V는 16과 240사이의 값을 가진다. YUV 모델은 흑백 텔레비전이 전국에 보유 중이던 시기에, 컬러 텔레비전의 개발로 고안되었다. 명암을 알려주는 휘도신호만 이용한 흑백 텔레비전의 색 공간 정보를 컬러로 불러오기 위하여 색차신호인 U, V를 추가하는 형태로 발전하였다. 따라서 무채색을 나타낼 때는 휘도신호 Y만 이용하면 된다. YUV 모델은 세 가지 신호로 색을 표현하는 RGB 모델 방식보다 색상을 세분화하지 못하여 영상의 화질이 저하될 수 있다는 단점이 있지만 적은 정보로 다양한 색상을 표현할 수 있어 보다 적은 양의 정보로 RGB 모델과 비슷한 화질을 나타낼 수 있다는 장점이 있다(Lee., 2003).

[그림 1][1]은 원본 이미지(Barns grand tetons)를 RGB 모델의 각 채널과 YUV 모델의 각 채널로 나타낸 그림이다. [그림 1]을 보면, 휘도신호 Y채널만으로 컬러 이미지가 흑백 이미지로 변환됨을 알 수 있다.

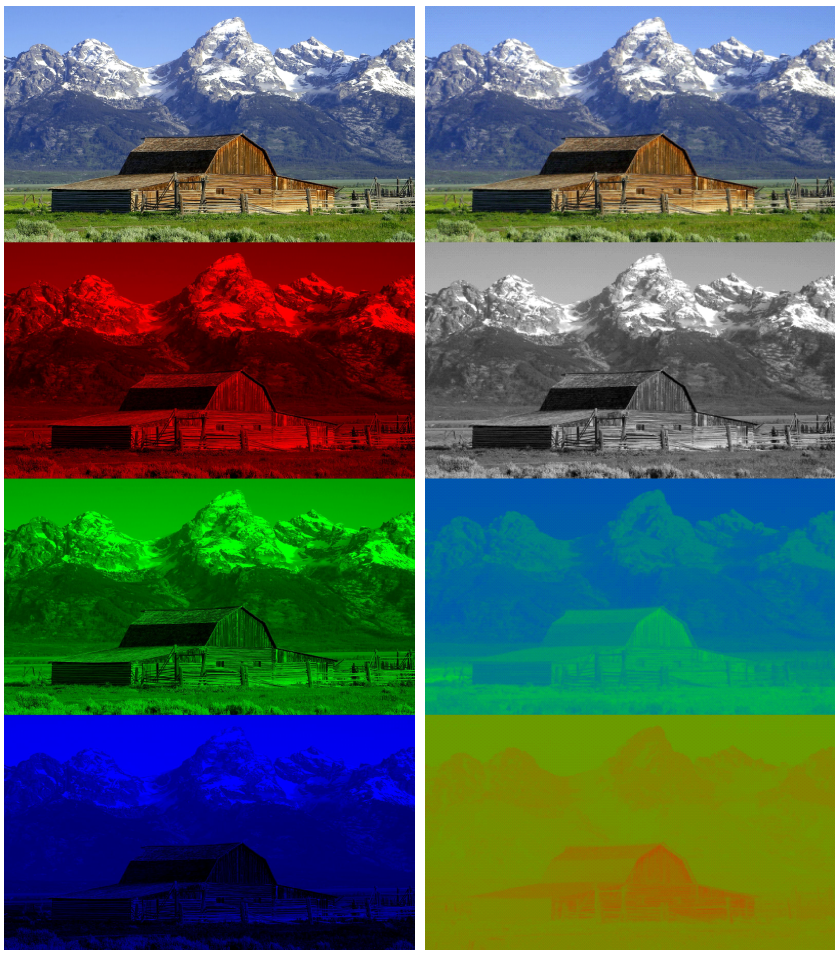
현재 대부분의 출력장치에서는 RGB 모델 방식으로 이미지의 색을 표현하고 있기 때문에 YUV 신호를 출력하기 위하여 색을 변환하는 색 공간 변환기(color space converter)가 만들어졌다. RGB 모델과 YUV 모델 사이에는 서로 변환할 수 있는 방법이 존재한다. 변환방법은 식 (2.1.1)과 같다.

---

1) Barn grand tetons, Nopira, Brianski, 2007,

$$
\begin{aligned}
Y &= 0.3R + 0.59G + 0.11B, \\
U &= (B - Y) \times 0.493, \\
V &= (R - Y) \times 0.877.
\end{aligned}
\tag{2.1.1}
$$

본 논문에서는 R의 readJPEG 함수를 이용하여 이미지 픽셀값을 출력하였다. readJPEG 함수는 R, G, B 3차원의 값을 0과 1사이의 실수로 변환한다. 변환된 R, G, B 차원을 YUV 모델의 Y로 변환하여 데이터로 이용하였다.

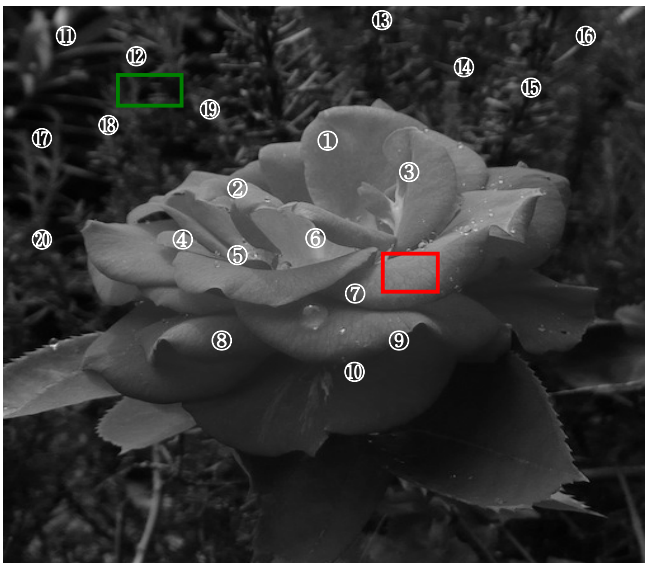


[그림 1] 이미지의 RGB 모델 방식 변환과 YUV 모델 방식 변환

2.2. 집단, 훈련 자료, 검증자료

본 절에서는 집단(class), 훈련 자료(training data), 검증자료(test point)에 대해 설명하고자 한다. 주어진 이미지에서 분류에 관심이 있는 영역을 지정하여 이를 집단이라고 한다. 각 집단을 대표할 수 있는 영역을 지정하여 훈련자료를 생성하고, 분류하고자 하는 새로운 픽셀 값을 검증자료라고 명명한다(Ghimire et al., 2012).

[그림 2][2]의 장미그림을 예시로 들어 설명하면 다음과 같다. 이미지 분류에 관심 있는 영역을 꽃잎(flower)과 풀잎(leaf)이라 하자. 꽃잎 집단을 대표하는 영역을 붉은색 박스로, 풀잎 집단을 대표하는 영역을 초록색 박스로 훈련자료의 영역을 표시한다. 검증자료 20개를 뽑아 각 검증자료가 꽃잎 집단에 속하는지 풀잎 집단에 속하는지를 분류하고자 한다.



[그림 2] 집단 및 훈련자료와 검증자료

본 논문에서는 분류에 관심 있는 집단으로 새로운 검증자료를 분류

2) lilytana, Do not cry, my rose: you are so pretty that I may love you!, 2008

하기 위해 LA 방법, GW 방법, 그리고 위치모수와 척도모수의 동시검정법인 K-S검정, CVM검정, Lepage검정, Cucconi검정을 이용하여 검증자료를 분류하고, 각각의 정분류율을 확인하여 어떤 방법이 분류에 효과적인지를 확인하고자 한다.

제 3 장 이미지 픽셀 분류 방법

이 장에서는 위치모수만을 고려한 가설검정 기반 분류법인 LA 방법과 GW 방법에 대해 설명하고, 위치모수와 척도모수의 동시검정법인 K-S 검정, CVM 검정, Lepage 검정, Cucconi 검정에 대해 소개하고자 한다.

3.1. Liao & Akritas 방법

Liao & Akritas는 가설검정을 기반으로 한 새로운 비모수적 분류방법을 제안하였다.

모평균이 μ_1, μ_2인 두 개의 집단 C_1, C_2가 존재하며, X는 입력변수(input variable)이다. 또한, 집단 C_1을 대표하는 훈련자료는 $(X_{11}, X_{12}, \cdots, X_{1m})$, 집단 C_2를 대표하는 훈련자료는 $(X_{21}, X_{22}, \cdots, X_{2n})$ 라고 하자. 여기서, m, n은 각 집단의 크기이다. LA 방법의 관심은 새로운 관측치 X_0가 집단 C_1과 집단 C_2 중 어느 집단으로 속하는 가이다.

LA 방법의 기본적인 아이디어는 만약 새로운 관측값 X_0가 집단 C_1에 속하는 경우라면, X_0을 집단 C_2로 분류했을 때, 두 집단 간 평균의 차이가 작아진다는 것이다. 이 아이디어를 통해 다음의 두 가지 검정을 설정할 수 있다. 첫 번째 검정을 T_1, 두 번째 검정을 T_2라고 하자.

T_1: X_0을 집단 C_1로 분류하여 새로운 집단 C_1'을 생성한다. 그리고 새로운 집단 C_1'과 집단 C_2를 평균 비교한다. 즉, 각 집단의 평균을 μ_1, μ_2라 할 때, $C_1' = (X_{0,}\, X_{11,}\, X_{12,} \cdots, X_{1m})$ 과 $C_2 = (X_{21,}\, X_{22,} \cdots, X_{2n})$을 귀무가설 $H_0 : \mu_1 = \mu_2$ 하에서 검정한다.

T_2: X_0을 집단 C_2로 분류하여 새로운 집단 C_2'을 생성한다. 그리고 집단 C_1과 새로운 집단 C_2'을 평균 비교한다. 즉, 각 집단의 평균을 μ_1, μ_2라 할 때, $C_1 = (X_{11,}\, X_{12,} \cdots, X_{1m})$ 과 $C_2' = (X_{0,}\, X_{21,}\, X_{22,} \cdots, X_{2n})$을 귀무가설 $H_0 : \mu_1 = \mu_2$ 하에서 검정한다.

LA 방법은 각 훈련자료의 크기에 따라 검정방법이 달라진다. 두 훈련자료의 크기가 같은 경우에는 t-검정을 이용한다. 첫 번째 검정 T_1과 두 번째 검정 T_2에서의 통계량을 각각 t_1, t_2라 하면 다음과 같이 정의할 수 있다.

분산이 동일한 경우 t_1, t_2는 다음으로 정의할 수 있다.

$$t_1 = \frac{\frac{n\overline{X}_1 + X_0}{n+1} - \overline{X}_2}{\sqrt{S_p\left(\frac{1}{n+1} + \frac{1}{n}\right)}},\ t_2 = \frac{\overline{X}_1 - \frac{n\overline{X}_2 + X_0}{n+1}}{\sqrt{S_p\left(\frac{1}{n} + \frac{1}{n+1}\right)}}, \qquad (3.1.1)$$

$$S_p = \frac{(n-1)s_x^2 + (n-1)s_y^2}{n+n-2} = \frac{s_x^2 + s_y^2}{2}.$$

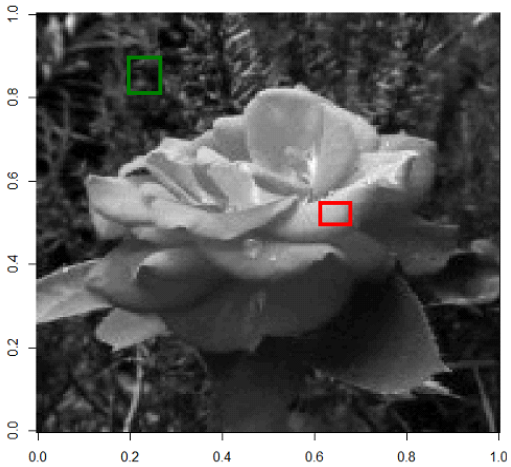
분산이 동일하지 않은 경우 t_1, t_2는 다음으로 정의할 수 있다.

$$t_1 = \frac{\frac{m\overline{X}_1 + X_0}{m+1} - \overline{X}_2}{\sqrt{\frac{s_x^2 + (X_0 - \overline{X}_1)^2}{m+1} + \frac{s_y^2}{n}}}, t_1 = \frac{\overline{X}_1 - \frac{n\overline{X}_2 + X_0}{n+1}}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2 + (X_0 - \overline{X}_1)^2}{n+1}}}. \quad (3.1.2)$$

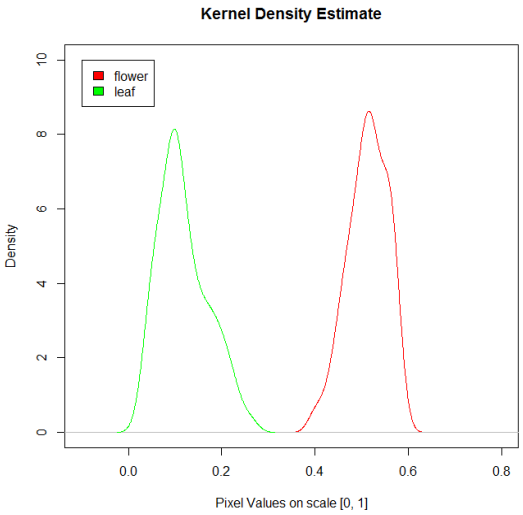
두 훈련자료의 크기가 다른 경우, $\overline{X}_1 > \overline{X}_2$ 이면, $X_0 > c_1\overline{X}_1 + c_2\overline{X}_2$이면 새로운 검증자료 X_0를 집단 1로 분류한다.

$$c_1 = \frac{\sqrt{m(m+1)} - m\sqrt{n(n+1)} + \sqrt{m(m+1)}\,n}{\sqrt{m(m+1)} + \sqrt{n(n+1)}},$$
$$c_2 = \frac{\sqrt{n(n+1)} + m\sqrt{n(n+1)} - \sqrt{m(m+1)}\,n}{\sqrt{m(m+1)} + \sqrt{n(n+1)}}. \quad (3.1.3)$$

첫 번째 검정 T_1과 두 번째 검정 T_2의 p-값을 비교하여 p-값이 더 작은 쪽으로 집단을 할당한다. 즉, T_1의 p-값이 더 작으면 X_0를 집단 C_1으로, T_2의 p-값이 더 작으면 집단 C_2로 할당하는 것이다. 하지만 LA 방법은 T_1과 T_2에서의 p-값이 모두 매우 작을 때, 오분류율이 매우 높다. 이를 [그림 1]의 장미그림을 가지고 간단히 확인해보자.



[그림 3] 좌표화한 [그림 1]의 훈련자료



[그림 4] 두 집단의 커널밀도함수

[그림 3]은 [그림 1]을 0과 1 사이의 값을 가지는 차원 R, G, B의 픽셀값으로 저장한 뒤, YUV 방식으로 변환하여 좌표위에 출력한 그림이다. 꽃잎 집단과 풀잎 집단의 훈련자료는 각각 붉은색, 초록색으로 표시하였다.

[그림 4]는 꽃잎 집단과 풀잎 집단의 커널밀도함수이다. 커널밀도함수를 살펴보면, 두 집단의 평균은 꽃잎 집단이 풀잎 집단에 비해 더 크고, 두 집단의 분산은 크게 차이가 나지 않아 보인다. 또한 꽃잎집단의 최소값이 풀잎 집단의 최대값보다 큰 것으로 보인다. 두 집단의 다섯수치요약, 평균, 분산의 결과는 [표 1]과 같다.

[표 1] 꽃잎 집단과 풀잎 집단의 다섯수치요약과 평균, 분산

	최소값	Q_1	중위수	Q_3	최대값	평균	분산
꽃잎	0.3979	0.4863	0.5153	0.5481	0.5843	0.5128	0.0018
풀잎	0.0278	0.0782	0.1061	0.1533	0.2631	0.1181	0.0028

Q_1: 제 1 사분위수, Q_3: 제 3 사분위수

훈련 자료가 아닌 이미지의 다른 영역에서 풀잎 집단과 꽃잎 집단의 검증자료를 각각 10개씩 총 20개를 랜덤으로 추출하였다. 검증자료 1~10은 꽃잎 집단, 11~20은 풀잎 집단이다. 이를 바탕으로 검정 T_1과 T_2를 수행하였다. 각 검정의 p-값이 매우 작아서 0으로 출력되었기 때문에 분류기준을 T 통계량으로 두어 분류를 진행하였다. 검정결과를 [표 2]로 정리하였다. T_1의 검정통계량 T_{LA1}이 T_2의 검정통계량 T_{LA2}보다 항상 큰 것을 볼 수 있다. 따라서 모든 검증자료가 꽃잎 집단으로 분류되었다. 이 결과로 Ghimire & Wang(2007)이 발견한 LA 방법의 단점을 확인할 수 있었다.

본 연구에서도 모의실험 시 모든 p-값이 매우 작아 거의 0에 가까운 값으로 계산어 비교가 어렵기 때문에 T 통계량을 이용하여 그 값이 큰 쪽으로 집단을 할당하였다.

[표 2] LA 방법을 이용한 [그림 3]의 이미지 픽셀 분류

TP	LA	T_{LA1}	T_{LA2}	value
1	Class 1	1205.516	1472.082	0.396667
2	Class 1	1154.753	1415.37	0.307765
3	Class 1	1229.286	1496.305	0.47302
4	Class 1	1232.744	1498.851	0.49949
5	Class 1	1106.802	1360.197	0.244745
6	Class 1	1232.754	1498.855	0.499608
7	Class 1	1159.386	1420.643	0.314549
8	Class 1	1232.456	1498.704	0.496235
9	Class 1	1233.666	1498.741	0.519059
10	Class 1	1233.075	1498.974	0.503922
11	Class 1	1046.857	1290.237	0.176353
12	Class 1	1065.229	1311.764	0.196549
13	Class 1	1029.592	1269.95	0.157804
14	Class 1	1032.499	1273.37	0.160902
15	Class 1	1038.833	1280.815	0.167686
16	Class 1	1007.703	1244.16	0.134745
17	Class 1	906.5385	1124.145	0.031412
18	Class 1	1023.123	1262.336	0.150941
19	Class 1	968.4989	1197.796	0.094314
20	Class 1	984.0738	1216.24	0.110275

Class 1: 꽃잎 집단, TP: 검증자료(test point), T_{LA1} : 검정 T_1의 통계량
Class 2: 풀잎 집단, LA: LA 방법의 분류결과, T_{LA2} : 검정 T_2의 통계량
value: 검증자료의 이미지픽셀값

3.2. Ghimire & Wang 방법

Ghimire & Wang(2012)은 앞선 예제에서 살펴보았듯이, 검정기반 분류방법인 LA 방법이 검정 T_1과 T_2의 p-값이 매우 작을 경우, 오분류율이 높다는 문제점을 발견하였다. 이에 LA 방법을 보완하여 새로운 이미지픽셀 분류방법을 고안하였다. GW 방법은 아주 작은 p-값의 문제를 완화하여 정분류율을 높인다.

PV_1을 검정 T_1에서의 p-값, PV_2를 검정 T_2에서의 p-값이라고 할 때, GW 방법은 아주 작은 임계값 ϵ(예를 들면, 0.001)을 두어 LA 방법을 다음과 같이 수정하였다.

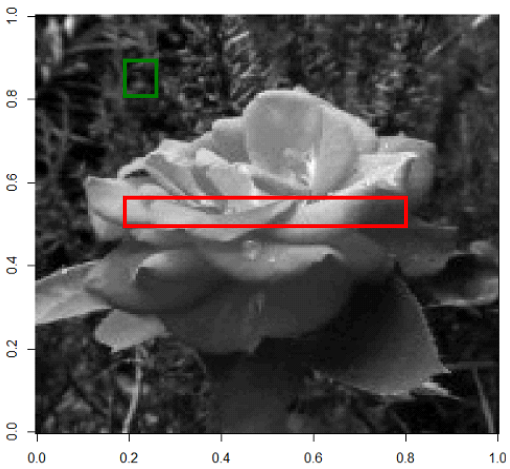
- $\max(PV_1, PV_2) \geq \epsilon$ 이면, 검증자료 X_0를 LA 방법으로 분류한다.
- $\max(PV_1, PV_2) < \epsilon$ 이면, 검증자료 X_0와 각 집단 C_1, C_2와의 거리가 가까운 쪽으로 분류한다.

즉, PV_1와 PV_2중 어느 하나라도 임계값 ϵ보다 크고, $PV_1 < PV_2$ 이면 검증자료 X_0를 집단 C_1로 분류하고, $PV_1 > PV_2$이면 X_0를 집단 C_2로 분류한다.

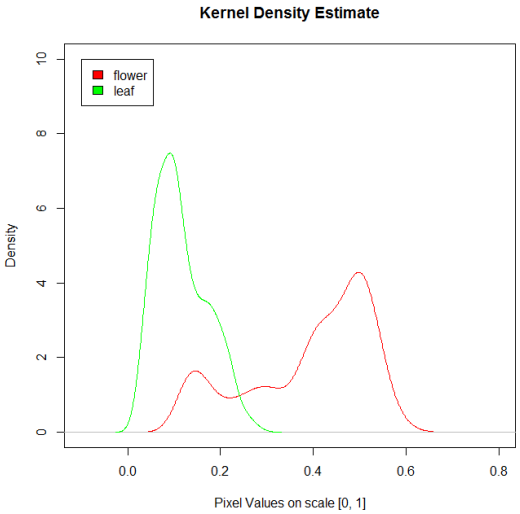
또한, PV_1와 PV_2 모두 임계값 ϵ보다 작은 경우, 검증자료 X_0가 집단 C_1과 가깝다면 집단 C_1로 분류하고, 집단 C_2와 가깝다면 집단 C_2로 분류한다. 점과 군집간의 거리를 구하는 다양한 방법들이 있지만, GW 방법은 단순하게 각 집단의 평균과 새로운 검증자료 X_0간의 거리 D를 비교하여 분류하는 방법을 선택하였다.

하지만 GW 방법도 평균만을 고려하였기 때문에 분산이 매우 커서 두 집단이 겹치는 상황에 대해서는 제대로 분류하지 못함을 발견하였

다. 이를 장미그림을 이용하여 간단히 확인해보자.



[그림 5] 좌표화한 [그림 1]의 훈련자료



[그림 6] 두 집단의 커널밀도함수

[그림 5]는 [그림 3]의 장미그림에서 한 집단의 분산을 크게 만들기 위해 꽃잎 집단의 훈련자료를 변경하였다. 꽃잎 집단과 풀잎 집단의 훈련자료는 각각 붉은색, 초록색으로 표시하였다.

[그림 6]은 꽃잎 집단과 풀잎 집단의 커널밀도함수이다. 커널밀도함수를 살펴보면, 꽃잎 집단의 평균과 분산이 풀잎 집단의 평균과 분산에 비해 더 크다. 또한 꽃잎집단의 픽셀값이 풀잎 집단의 픽셀값과 중복되는 부분도 존재한다. 두 집단의 다섯수치요약, 평균, 분산의 결과는 [표 3]과 같다.

[표 3] 꽃잎 집단과 풀잎 집단의 다섯수치요약과 평균, 분산

	최소값	Q_1	중위수	Q_3	최대값	평균	분산
꽃잎	0.1186	0.2987	0.4292	0.4998	0.6224	0.3922	0.0175
풀잎	0.0209	0.0717	0.1040	0.1556	0.2834	0.1156	0.0032

Q_1: 제 1 사분위수, Q_3: 제 3 사분위수

훈련 자료가 아닌 이미지의 다른 영역에서 풀잎 집단과 꽃잎 집단의 검증자료를 각각 10개씩 랜덤으로 추출하였다. 검증자료 1~10은 꽃잎 집단, 11~20은 풀잎 집단이다. 이를 바탕으로 검정 T_1과 T_2를 수행하였다. 각 검정의 p-값이 매우 작아서 아주 작은 임계값 $\epsilon(0.001)$보다 작으므로 분류기준을 새로운 검증자료 X_0와 각 집단 평균의 거리 D를 이용하여 분류를 진행하였다. 검정결과를 [표 4]로 정리하였다. 풀잎 집단에서는 오분류가 하나도 일어나지 않았지만 분산이 큰 꽃잎집단에서는 50%(5개, TP=1, 2, 4, 7, 9)의 오분류가 발생했다. 오분류가 일어난 검증자료 1, 2, 4, 7, 9의 값은 각각 0.201059, 0.211922, 0.252196, 0.20898, 0.169765으로 두 집단의 커널밀도함수가 겹쳐 있는 부분에 속하는 값이다. 이 결과로 GW 방법도 분산을 고려하지 않아 오분류율이 높다는 것을 확인하였다.

[표 4] GW 방법을 이용한 [그림 3]의 이미지 픽셀 분류

TP	GW	D_1	D_2	value
1	Class 2	0.191122	0.085417	0.201059
2	Class 2	0.180259	0.096279	0.211922
3	Class 1	0.038769	0.23777	0.353412
4	Class 2	0.139985	0.136554	0.252196
5	Class 1	0.009741	0.286279	0.401922
6	Class 1	0.108691	0.167848	0.28349
7	Class 2	0.183201	0.093338	0.20898
8	Class 1	0.135514	0.141025	0.256667
9	Class 2	0.222416	0.054123	0.169765
10	Class 1	0.012642	0.289181	0.404824
11	Class 2	0.211867	0.064672	0.180314
12	Class 2	0.247279	0.02926	0.144902
13	Class 2	0.245789	0.03075	0.146392
14	Class 2	0.278416	0.001877	0.113765
15	Class 2	0.278887	0.002348	0.113294
16	Class 2	0.204024	0.072515	0.188157
17	Class 2	0.211671	0.064868	0.18051
18	Class 2	0.285632	0.009093	0.106549
19	Class 2	0.191985	0.084554	0.200196
20	Class 2	0.283475	0.006936	0.108706

Class 1: 꽃잎 집단, TP: 검증자료(test point), D_1 : 검정 T_1의 통계량

Class 2: 풀잎 집단, GW: GW 방법의 분류결과,D_2 : 검정 T_2의 통계량

value: 검증자료의 이미지픽셀값

3.3. 위치모수와 척도모수의 동시검정법

앞 절에서 간단한 예제를 통하여 가설검정을 기반으로 한 LA 방법과 GW 방법은 p-값이 너무 작거나 한 집단의 분산이 매우 크면, 이미지 픽셀 분류 시 오분류율이 높다는 것을 확인하였다. 가설검정을 기반으로 하되, 위치모수와 척도모수의 동시검정법을 이용하여 이미지를 분류한다면 기존의 방법들을 보완하여 이미지 픽셀값의 정분류율이 높을 것이라 예상하였다.

본 절에서는 위치모수와 척도모수의 동시검정법에 대해 설명하고자 한다.

3.3.1. Kolmogorov - Smirnov 검정

K-S검정은 경험적 분포함수(empirical distribution function)의 차를 이용한 분포의 동일성 검정방법으로 두 모집단이 서로 동일한 분포를 가지는지를 확인하기 위한 것이 목적이다.

$X_1, X_2, \cdots, X_m$을 연속분포인 $F(x)$로부터 얻은 m개의 독립확률표본이라 하고, $Y_1, Y_2, \cdots, Y_n$을 연속분포인 $G(y)$로부터 얻은 n개의 독립확률표본이라 하자. K-S 검정의 귀무가설 H_0과 대립가설 H_1은 다음과 같다.

H_0 : 두 분포는 동일하다(모든 x에 대해 $F(x) = G(y)$).

H_1 : 두 분포는 동일하지 않다.

K-S 검정에서는 단측검정인지 양측검정인지에 따라 검정통계량을 다르게 정의한다. 두 분포가 동일하지 않은 경우는 총 세 가지로 볼 수

있다. 먼저, 적어도 한 점 x에 대해 $F(x) < G(y)$인 경우, 적어도 한 점 x에 대해 $F(x) > G(y)$인 경우, 적어도 한 점 x에 대해 $F(x) \neq G(y)$인 경우이다. 각각의 경우에 대해 검정통계량을 살펴보면 다음과 같다.

$$\begin{aligned} &1)\ F(x) < G(y) \quad D_1^+ = \sup\{G_n(y) - F_m(x)\} \\ &2)\ F(x) > G(y) \quad D_1^- = \sup\{F_m(x) - G_n(y)\} \\ &3)\ F(x) \neq G(y) \quad D_1 = \sup\{|F_m(x) - G_n(y)|\} = \max\{D_1^+, D_1^-\} \end{aligned} \tag{3.3.1}$$

3.3.2. Cramér - von Mises 검정

Cramér-von Mises 검정은 K-S검정과 마찬가지로 경험적 분포함수의 차를 이용한 방법으로, 경험적 분포함수의 차를 제곱하여 두 집단이 동일한 분포로부터 온 것인지 알아보고자 하는 검정이다.

$X_1, X_2, \cdots, X_m$을 연속분포인 $F(x)$로부터 얻은 m개의 독립확률표본이라 하고, $Y_1, Y_2, \cdots, Y_n$을 연속분포인 $G(y)$로부터 얻은 n개의 독립확률표본이라 하자. CVM 검정의 귀무가설 H_0과 대립가설 H_1은 다음과 같다.

H_0 : 두 분포는 동일하다(모든 x에 대해 $F(x) = G(y)$).

H_1 : 두 분포는 동일하지 않다.

r_i, s_j를 각각 혼합표본에서 집단 X의 순위, 집단 Y의 순위라고 하면 Cramér - von Mises의 검정통계량 C는 다음과 같이 정의된다.

$$C = \frac{S}{mn(m+n)} - \frac{4mn-1}{6(m+n)},$$
$$S = m\sum_{i=1}^{m}(r_i - i)^2 + n\sum_{j=1}^{n}(s_j - j)^2. \quad (3.3.2)$$

Cramér - von Mises의 검정통계량 C가 클수록 귀무가설 H_0를 기각한다.

3.3.3. Lepage 검정

Lepage 검정은 Lepage(1971)에 의해 제안되었으며, 위치모수와 척도모수의 동시검정법 중 가장 널리 알려진 방법이다(Marco, 2012). Lepage 검정은 이표본 위치문제 검정법인 윌콕슨 순위합검정(Wilcoxon rank sum test)과 이표본 척도문제 검정법인 앤서리-브래들리 검정(Ansari-Bradley test)을 함께 고려한 검정방법이다. 먼저 윌콕슨의 순위합검정과 앤서리-브래들리 검정을 확인하고, Lepage 방법에 대해 알아보자.

1) 윌콕슨 순위합 검정(Wilcoxon rank sum test)

윌콕슨 순위합 검정(Wilcoxon rank sum test)은 이표본 위치문제에서 가장 많이 사용되는 비모수적 방법이다. 이 검정방법은 집단의 혼합표본에서 각 관측값이 작은 값부터 오름차순으로 순위를 부여한 뒤, 집단의 순위합을 이용하는 방법이다.

순위합 검정의 귀무가설 H_0과 대립가설 H_1은 다음과 같다.

H_0 : 두 집단의 위치모수(location parameter)는 동일하다.

H_1 : 두 집단의 위치모수(location parameter)는 동일하지 않다.

$X_1, X_2, \cdots, X_m$을 연속분포인 $F(x)$로부터 얻은 m개의 독립확률표본(independent random samples)이라 하고, $Y_1, Y_2, \cdots, Y_n$을 연속분포인 $G(y)=F(ay+b)$, $a>0$로부터 얻은 n개의 독립확률표본이라 하자. 또한, 혼합표본에서 Y_i의 순위를 R_i라 하면 순위합 통계량 W는 다음과 같다.

$$W=\sum_{i=1}^{n} R_i . \tag{3.3.3}$$

각 확률표본의 크기 m, n이 충분히 클 때는, 중심극한정리를 이용하여 대표본 정규근사로 표준화된 W 통계량을 이용하여 검정할 수 있다.

2) 앤서리-브래들리 검정(Ansari-Bradley test)

윌콕슨 순위합 검정에서는 위치모수(location parameter)에 대한 문제를 검정했다면, 앤서리-브래들리 검정(Ansari-Bradley test)은 두 집단의 위치모수가 같을 때, 혼합표본이 순위에 영향을 주는 것이 척도모수(scale parameter)라는 발상에서 착안하였다. 척도모수를 고려하는 여러 종류의 분포무관 검정법 중 대표적인 것이 바로 앤서리-브래들리 검정이다. 이 검정방법은 집단의 혼합표본에서 관측값이 작은 값부터 순위를 부여한 뒤, 앤서리-브래들리 순위(Ansari-Bradley ranking)를 통해 집단의 앤서리-브래들리 스코어(Ansari-Bradley score)를 이용하는 방법이다. 앤서리-브래들리 순위는 혼합표본의 크기가 홀수인지, 짝수인지에 따라 다르게 부여되며, 중심으로 갈수록 스코어가 높아진다.

앤서리-브래들리 검정의 귀무가설 H_0과 대립가설 H_1은 다음과 같다.

H_0 : 두 집단의 척도모수(scale parameter)는 동일하다.

H_1 : 두 집단의 척도모수(scale parameter)는 동일하지 않다.

$X_1, X_2, \cdots, X_m$을 연속분포인 $F(u)$로부터 얻은 m개의 독립확률표본(independent random samples)이라 하고, $Y_1, Y_2, \cdots, Y_n$을 연속분포인 $G(u)=F(\theta u)$로부터 얻은 n개의 독립확률표본이라 하자. θ는 척도모수이다. 또한, 혼합표본에서 X_i의 순위를 RS_i라 하고 앤서리-브래들리 스코어를 $AB(RS_i)$라 한다.

[표 5] N이 홀수인 경우($N=m+n$)

RS_i	1	2	$\cdots$	$\frac{N+1}{2}$	$\cdots$	$N-1$	N
$AB(RS_i)$	1	2	$\cdots$	$\frac{N+1}{2}$	$\cdots$	2	1

[표 6] N이 짝수인 경우($N=m+n$)

RS_i	1	2	$\cdots$	$\frac{N}{2}$	$\frac{N}{2}+1$	$\cdots$	$N-1$	N
$AB(RS_i)$	1	2	$\cdots$	$\frac{N}{2}$	$\frac{N}{2}$	$\cdots$	2	1

앤서리-브래들리 통계량 A는 다음과 같다.

$$A = \sum_{i=1}^{m} AB(RS_i). \tag{3.3.4}$$

각 확률표본의 크기 m, n이 충분히 클 때는, 중심극한정리를 이용하여 대표본 정규근사로 표준화된 A 통계량을 이용하여 검정할 수 있다.

Lepage 검정은 앞서 소개한 윌콕슨 순위합검정과 앤서리-브래들리 검정을 동시에 고려하여 만든 새로운 위치모수와 척도모수의 동시검정법(location-scale test)이다. 표준화된 윌콕슨 순위합 통계량과 앤서리-브래들리 통계량의 합으로 통계량을 고안하였다.

$X_1, X_2, \cdots, X_m$을 연속분포인 $F(x)$로부터 얻은 m개의 독립확률표본이라 하고, $Y_1, Y_2, \cdots, Y_n$을 연속분포인 $G(y) = F(ay+b),\ a>0$로부터 얻은 n개의 독립확률표본이라 하자. Lepage 검정의 귀무가설 H_0과 대립가설 H_1은 다음과 같다.

H_0 : 두 분포는 동일하다(즉, $a=1,\ b=0$).

H_1 : 두 분포는 동일하지 않다(즉, $a \neq 1,\ b \neq 0$).

윌콕슨 순위합 통계량 W와 앤서리-브래들리 통계량 A를 이용한 Lepage 검정통계량 L은 다음과 같이 정의된다.

$$L=\left(\frac{W-\mu_W}{\sigma_W}\right)^2+\left(\frac{A-\mu_A}{\sigma_A}\right)^2, \tag{3.3.5}$$

여기서, 귀무가설 하의 평균과 분산인 μ_W와 μ_A, σ_W와 σ_A는 다음과 같다.

$$\mu_W=E(W|H_0) \quad =\frac{1}{2}m(N+1),$$

$$\sigma_W^2=VAR(W|H_0)=\frac{1}{12}mn(N+1),$$

$$\mu_A=E(A|H_0) \quad =\begin{cases}\frac{m(N+1)}{4} & ,N\text{이 짝수}\\ \frac{m(N+1)^2}{4N} & ,N\text{이 홀수},\end{cases}$$

$$\sigma_A^2=VAR(A|H_0)=\begin{cases}\frac{mn(N+2)(N-2)}{48(N-1)} & ,N\text{이 짝수}\\ \frac{mn(N+1)(N^2+3)}{48N^2} & ,N\text{이 홀수}.\end{cases}$$

각 확률표본의 크기 m, n이 충분히 클 때, Lepage 통계량 L은 윌콕슨의 순위합 통계량 W와 앤서리-브래들리 통계량 A가 서로 귀무가설 H_0하에서 서로 상관이 없으며, 정규근사된 통계량의 제곱의 합으로 이루어졌기 때문에 자유도가 2인 χ^2분포로 수렴한다.

3.3.4. Cucconi 검정

Cucconi 검정은 Cucconi(1968)에 의해 제안되었으며, Leapge 검정과 마찬가지로 위치모수와 척도모수의 동시검정법이다. 이 방법은 이탈리아어로 발표되어 Lepage 검정보다는 상용화되지 못했지만, 어떤 경우에 있어서는 Lepage 검정보다 우수한 결과를 보인다(Marco, 2009).

$X=(X_1, X_2, \cdots, X_m)$ 과 $Y=(Y_1, Y_2, \cdots, Y_n)$ 가 각각 독립적인 확률표본이라고 하자. Cucconi 검정의 귀무가설 H_0과 대립가설 H_1은 다음과 같다.

H_0 : 두 분포는 동일하다.

H_1 : 두 분포는 동일하지 않다.

Cucconi 검정통계량은 다음과 같이 정의된다.

$$C=\frac{U^2+V^2-2\rho UV}{2(1-\rho^2)}, \tag{3.3.6}$$

여기서, U와 V, 그리고 ρ는 다음과 같다.

$$U=\frac{6\sum_{i=1}^{m}W_i^2-m(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)(8N+11)/5}},$$

$$V=\frac{6\sum_{i=1}^{m}(N+1-W_i)^2-m(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)(8N+11)/5}},$$

$$\rho=\frac{2(N^2-4)}{(2N+1)(8N+11)}-1.$$

U는 집단 X의 순위제곱합 W_i를 기초로 구성한 반면, V는 집단 X의 역순위(counter-ranks)제곱합 $N+1-W_i$를 기초로 구성하였다. Cucconi(1968)는 귀무가설 H_0하에서 $E(U)=E(V)=0$, $VAR(U)=VAR(V)=1$임을 증명하였다. 여기서 U와 V는 음의 관계이다(negative dependent). $\rho=CORR(U,V)=COVAR(U,V)$이며, $N>2$이면, ρ의 범위는 $-1<\rho<-7/8$ 이다.

귀무가설 H_0하에서 $\rho_0=-7/8$일 때, U와 V의 점근적 확률밀도함수(asymptotic probability density function)는 다음의 이변량 정규분포를 따른다.

$$f(u,v)=\frac{1}{2\pi\sqrt{1-\rho_0^2}}\exp\left(-\frac{u^2+v^2-2\rho_0uv}{2(1-\rho_0^2)}\right). \qquad (3.3.7)$$

제 4 장 모의실험 및 결과

본 장에서는 3장에서 소개한 LA 방법, GW방법, 그리고 위치모수와 척도모수의 동시검정법인 K-S 검정, CVM 검정, Lepage 검정, Cucconi 검정, 총 여섯 가지 방법을 이용하여 이미지 픽셀의 분류를 수행하였으며, 각 방법의 정분류율을 비교하여 어느 방법이 이미지 픽셀 분류를 가장 잘 하는지를 확인하였다.

4.1. 모의실험 방법

이 절에서는 모의실험의 과정을 설명하고, 모의실험에 이용한 집단에 대해 설명하고자 한다. 모의실험의 흐름도는 [그림 7]과 같다.

먼저 이미지 픽셀값을 0과 1사이의 값을 가지는 RGB 차원의 색공간으로 불러왔다. 이를 YUV 모델을 이용하여 하나의 차원으로 축소시킨 뒤, 각 집단을 대표하는 훈련자료를 설정하였다. 이는 [표 8]에 정리되어 있다. 그 이후, 각 집단별 검증자료를 10개씩, 총 30개의 검증자료를 추출하였다.

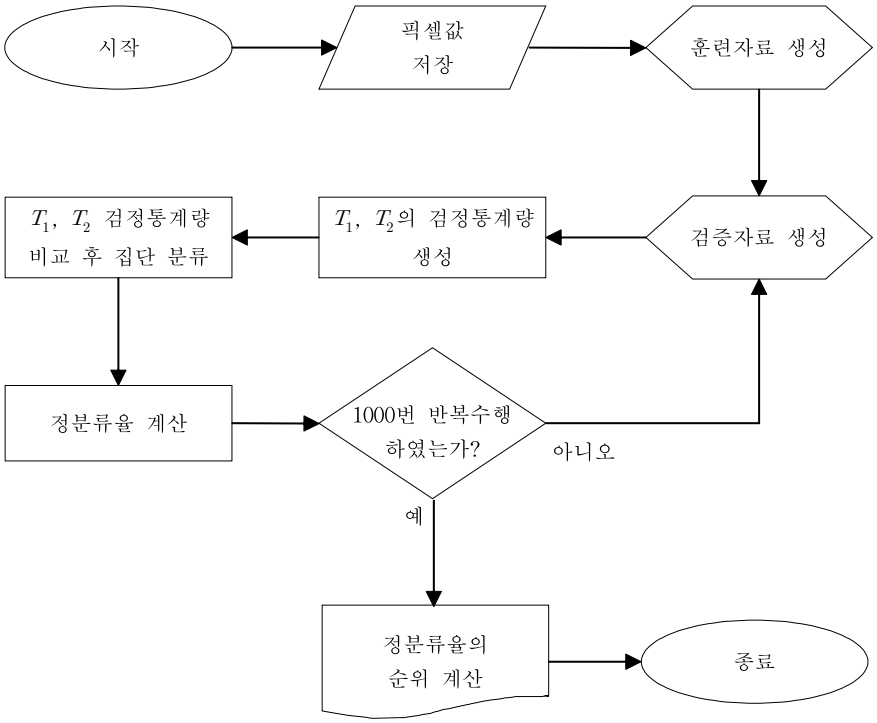
모의실험은 [표 7]과 같이 총 두 가지로 나누어 진행하였다.

[표 7] 모의실험 설계

	모의실험 설계 1	모의실험 설계 2
시나리오	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \ll \sigma_2^2$
실험 집단	집단 1: 산 집단 2: 호수	집단 1: 산 집단 2: 나무

첫 번째는 자료의 분산이 비슷하여 두 집단의 픽셀값이 겹치지 않는 경우로, 비교적 명확한 픽셀값의 차이를 보인다. 산 집단과 호수 집단을 이용하여 모의실험을 진행하였다.

두 번째는 한 집단의 분산이 매우 커서 두 집단의 픽셀값들이 서로 겹치는 부분이 존재하는 경우이다. 두 번째 상황에서는 산 집단과 숲 집단을 이용하였다.



[그림 7] 모의실험 흐름도

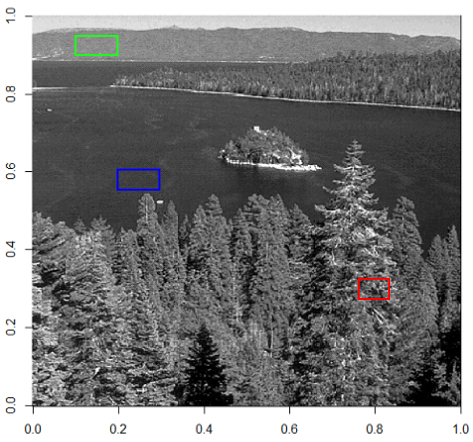
각 실험마다 다음 과정을 1000번 반복시행 하였다.

1. 여섯 가지 검정 방법으로 첫 번째 검정 T_1, 두 번째 검정 T_2의 검정통계량 생성(LA, GW, Lepage, Cucconi, K-S, CVM)
2. 검정 T_1에서 얻은 검정통계량와 검정 T_2에서 얻은 검정통계량을 비교하여 새로운 검증자료 X_0를 분류
3. 정분류율 계산
4. 각 모의실험 정분류율의 평균으로 순위 측정

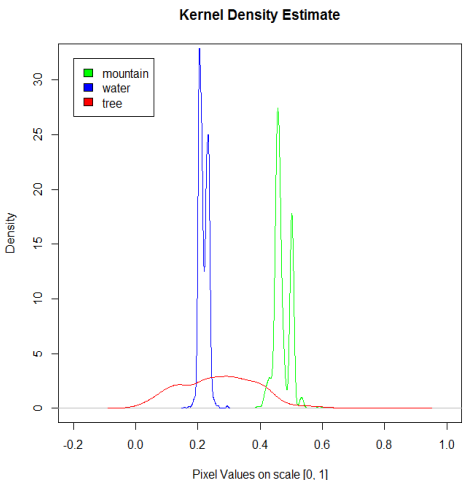
반복시행을 통하여 얻어진 정분류율의 평균을 계산하여 여섯 가지 방법을 비교하고, 어느 방법의 정분류율이 높은지 살펴보았다.

[표 8] 각 집단 훈련자료의 정보

	X	Y	차원	크기
산	(0.10, 0.20)	(0.90, 0.95)	52*26	1352
호수	(0.20, 0.30)	(0.55, 0.60)	53*26	1378
나무	(0.78, 0.82)	(0.25, 0.30)	22*27	594

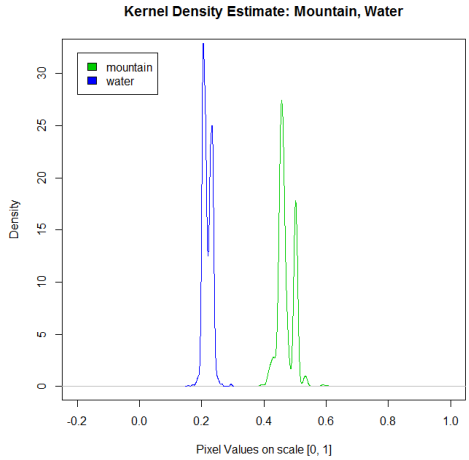


[그림 8] 좌표화한 이미지와 훈련자료

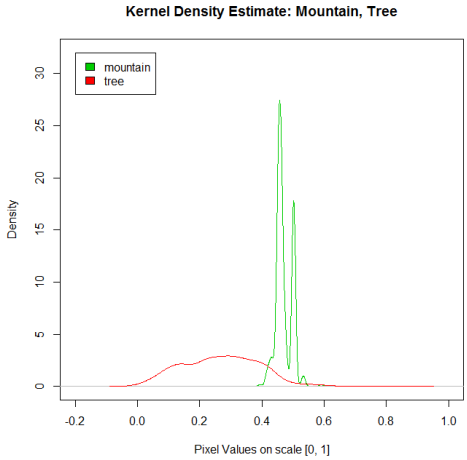


[그림 9] 세 집단의 커널밀도함수

[그림8][3]은 Bolshoy Vagilsky 호수 이미지이다. 각 집단은 산(집단 1), 호수(집단 2), 나무(집단 3)이며, 집단의 훈련자료는 각각 초록색, 파란색, 붉은색으로 표시하였다. [그림 9], [그림 10], [그림 11]은 각 집단의 커널밀도함수이다.



[그림 10] 커널밀도함수: 산, 호수



[그림 11] 커널밀도함수: 산, 나무

먼저 데이터 특징을 살펴보기 위하여, 세 집단의 다섯수치요약, 평균, 분산의 결과를 살펴보았다.

[표 9] 세 집단(산, 호수, 나무)의 다섯수치요약과 평균, 분산

	최소값	Q_1	중위수	Q_3	최대값	평균	분산	TP
산	0.3922	0.4549	0.4627	0.4980	0.6039	0.4701	0.0007	1~10
호수	0.1569	0.2078	0.2157	0.2314	0.2941	0.2182	0.0002	11~20
나무	0.2692	0.1618	0.2706	0.3598	0.8588	0.2692	0.0153	21~30

Q_1: 제 1 사분위수, Q_3: 제 3 사분위수, TP: 각 집단에 속하는 검증자료의 번호

3) http://decsai.ugr.es/cvg/CG/base.htm

[표 9]의 결과를 살펴보면, 산의 훈련자료와 호수의 훈련자료는 평균이 각각 0.4701, 0.2182로 [그림 9]와 [그림 10]의 커널 밀도함수에서도 살펴 볼 있듯이 서로 크게 차이가 있고, 산 훈련자료의 모든 픽셀 값이 호수 훈련자료의 모든 픽셀 값보다 큰 값을 가지고 있으므로 산과 호수의 픽셀값은 서로 겹치지 않음을 알 수 있다. 또한, 나무의 훈련자료는 다른 두 집단의 훈련자료보다 분산이 월등히 커서, 이미지 픽셀값이 전반적으로 0부터 1사이에 넓게 퍼져있음을 확인하였다.

이러한 정보를 통하여 산과 호수의 비교는 위치모수만을 고려한 GW 방법이 가장 정분류율이 높을 것으로 예상되며, 산과 나무의 비교는 위치모수와 척도모수의 동시검정법이 높을 것으로 예상된다.

4.2. 모의실험 결과

[그림 7]의 절차로 진행한 모의실험의 결과를 살펴보자. 우선 [표 10]에는 산 집단과 호수 집단의 1000번째 모의실험 결과를, [표 11]에는 산 집단과 호수 집단의 1000번째 모의실험결과를 실었다. 1000번째 모의실험에서, 산 집단과 호수집단을 비교한 경우, LA 방법은 검증자료를 산 집단(집단1)로 분류하였다. 반면 GW 방법, K-S 검정, CVM 검정, Cucconi 검정은 산 집단에 속하는 1~10번의 검증자료를 산 집단으로, 호수집단에 속하는 11~20번의 검증자료를 호수 집단으로 제대로 분류하였다. 산 집단과 나무 집단을 비교한 경우, 모든 방법에서 오분류가 존재함을 알 수 있다. 1000번의 모의실험의 정분류율을 계산하여 더 자세히 확인해보자.

[표 12]는 1000번 모의실험의 정분류율 중 일부만 나타낸 결과이다. 산 집단과 호수 집단을 분류하는 경우에, LA 방법을 제외한 다른 다섯 가지 방법의 정분류율이 높은 것으로 보인다. 산 집단과 나무 집단을

분류하는 경우에는 여섯 가지 방법 중 어떤 방법의 정분류율이 높은지 쉽게 보이지 않는다. 명확한 판단을 위하여 정분류율의 평균과 순위를 계산하였다.

[표 13]은 검증자료 1~10, 11~20, 21~30을 이용하여 계산한 정분류율 결과이다. 산 검증자료가 산 집단으로, 호수 검증자료가 호수 집단으로, 나무 검증자료가 나무 집단으로 제대로 분류되는지를 살펴보고자 하였다. 산 집단과 호수 집단을 비교하는 경우, LA 방법은 산 검증자료를 산 집단으로 분류하지 못하고 모든 검증자료를 호수 집단으로 분류한 반면, 위치모수만 고려한 GW방법과 위치모수와 척도모수의 동시검정법을 이용한 모든 분류는 모두 정분류율이 90% 이상으로 매우 높게 나타났다. 산 집단과 나무 집단을 비교하는 경우, 산 검증자료의 정분류율이 90% 이상인 방법은 LA 방법과 Cucconi 검정이었으며, 가장 정분류율이 낮은 방법은 CVM 검정이다. 나무 검증자료의 정분류율을 살펴보면, 위치모수와 척도모수의 동시검정법 중 K-S 검정, Lepage 검정, CVM 검정의 정분류율이 80%이상으로 높았으며, LA 방법과 GW 방법은 정분류율 순위 중 각각 5위, 6위로 낮은 순위를 보였다.

[표 14]에는 모의실험별 각 방법의 정분류율 평균과 순위를 실었다. 결과를 보면, 산 집단과 호수 집단을 분류하는 경우에는 GW 방법의 정분류율이 99.95%로 가장 우수하였으나, 위치모수와 척도모수의 동시검정법인 K-S 검정, CVM 검정, Lepage 검정, Cucconi 검정의 정분류율도 95%이상으로 매우 좋았다. 산 집단과 나무 집단의 분류에서는 LA 방법의 정분류율이 84.3%로 가장 높았으나 K-S 검정, Lepage 검정, Cucconi 검정방법의 정분류율 역시 80% 이상으로 크게 차이나지 않았다.

실험결과를 종합해보면, 분산이 같은 경우엔 GW방법의 정분류율이 높으며, 분산이 다른 경우에는 위치모수와 척도모수의 동시검정법 중 Cucconi 검정법을 이용한 이미지픽셀의 분류가 정분류율이 가장 높았다.

[표 10] 산 집단과 호수 집단의 비교: 1000번째 모의실험 분류결과

TP	LA	GW	K-S	CVM	Lepage	Cucconi	value
1	Class 2	Class 1	Class 1	Class 1	Class 2	Class 1	0.411765
2	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.427451
3	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.462745
4	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.458824
5	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.435294
6	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.45098
7	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.458824
8	Class 2	Class 1	Class 1	Class 1	Class 2	Class 1	0.411765
9	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.47451
10	Class 2	Class 1	Class 1	Class 1	Class 1	Class 1	0.454902
11	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.227451
12	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.223529
13	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.258824
14	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.211765
15	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.203922
16	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.235294
17	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.239216
18	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.203922
19	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.211765
20	Class 2	Class 2	Class 2	Class 2	Class 2	Class 2	0.25098

TP: 검증자료, value: 검증자료의 픽셀값, Class 1: 산 집단, Class 2: 호수 집단, value: 검증자료의 이미지픽셀값

[표 11] 산 집단과 나무 집단의 비교: 1000번째 모의실험 분류결과

TP	LA	GW	K-S	CVM	Lepage	Cucconi	value
1	Class 1	Class 3	Class 3	Class 3	Class 3	Class 3	0.411765
2	Class 1	Class 1	Class 1	Class 3	Class 1	Class 1	0.427451
3	Class 1	Class 1	Class 1	Class 1	Class 1	Class 1	0.462745
4	Class 1	Class 3	Class 1	Class 1	Class 1	Class 1	0.458824
5	Class 1	Class 1	Class 1	Class 3	Class 1	Class 1	0.435294
6	Class 1	Class 1	Class 1	Class 1	Class 1	Class 1	0.45098
7	Class 1	Class 3	Class 1	Class 1	Class 1	Class 1	0.458824
8	Class 1	Class 3	Class 3	Class 3	Class 3	Class 3	0.411765
9	Class 1	Class 1	Class 1	Class 1	Class 1	Class 1	0.474510
10	Class 1	Class 1	Class 1	Class 1	Class 1	Class 1	0.454902
11	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.219608
12	Class 1	Class 1	Class 1	Class 1	Class 1	Class 1	0.482353
13	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.235294
14	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.258824
15	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.329412
16	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.341177
17	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.266667
18	Class 3	Class 3	Class 3	Class 3	Class 3	Class 3	0.298039
19	Class 3	Class 1	Class 3	Class 3	Class 3	Class 3	0.380392
20	Class 1	Class 1	Class 1	Class 1	Class 1	Class 1	0.482353

TP: 검증자료, value: 검증자료의 픽셀값, Class 1: 산 집단, Class 3: 나무 집단, value: 검증자료의 이미지픽셀값

[표 12] 각 방법의 정분류율 일부

	산 vs. 호수						산 vs. 나무					
s	LA12	GW12	K-S12	CVM12	LP12	CU12	LA13	GW13	K-S13	CVM13	LP12	CU12
1	0.5	1	1	1	0.95	1	0.8	0.8	0.7	0.7	0.7	0.75
2	0.5	1	0.95	0.95	0.95	1	0.8	0.8	0.85	0.65	0.85	0.8
3	0.5	1	0.95	0.95	0.9	0.95	0.85	0.7	0.9	0.9	0.9	0.85
4	0.5	1	1	1	0.95	1	0.95	0.65	0.9	0.8	0.9	0.95
5	0.5	1	1	0.95	0.95	1	0.75	0.7	0.75	0.65	0.75	0.75
(중략)												
996	0.5	1	1	1	0.95	1	0.9	0.8	0.95	0.9	0.95	0.85
997	0.5	1	1	1	1	1	0.8	0.65	0.9	0.8	0.9	0.85
998	0.5	1	1	1	1	1	0.9	0.85	0.85	0.75	0.85	0.9
999	0.5	1	1	1	0.85	0.95	0.8	0.75	0.7	0.5	0.7	0.7
1000	0.5	1	1	1	0.9	1	0.9	0.65	0.8	0.7	0.8	0.8

s: 모의실험 차수, LA__: LA 방법 이용, GW__: GW 방법 이용,
LP__: Lepage 검정 이용, CU__: Cucconi 검정 이용, K-S__: Kolbogorov-Smirnov 검정 이용, CVM: Craimer-von Mise 검정 이용

[표 13] 검정 T_1과 T_2에서의 정분류율

	산 vs. 호수($m=10,\ n=20$)				산 vs. 나무($m=10,\ n=30$)			
검정법	C_1'vs.C_2	순위	C_1vs.C_2'	순위	C_1'vs.C_3	순위	C_1vs.C_3'	순위
LA	0	6	1	1	0.9634	1	0.7207	5
GW	0.9992	2	1	1	0.8595	5	0.6484	6
K-S	0.9818	3	0.9951	5	0.8620	3	0.8029	2
CVM	1	1	0.9799	6	0.6323	6	0.842	1
Lepage	0.9072	5	1	1	0.8620	3	0.8029	2
Cucconi	0.9689	4	1	1	0.9072	2	0.7588	4

$C_1 : (x_{1,} x_{2,} \cdots, x_m)$, $C_1' : (x_{0,} x_{1,} x_{2,} \cdots, x_m)$,
$C_2 : (x_{m+1,} x_{m+2,} \cdots, x_n)$, $C_2' : (x_{0,} x_{m+1,} x_{m+2,} \cdots, x_n)$,
$C_3 : (x_{2m+1,} x_{2m+2,} \cdots, x_n)$, $C_3' : (x_{0,} x_{2m+1,} x_{2m+2,} \cdots, x_n)$,

[표 14] 모의실험별 각 방법의 정분류율 평균과 순위

	산 vs. 호수						산 vs. 나무					
	LA12	GW12	K-S12	CVM12	LP12	CU12	LA13	GW13	K-S13	CVM13	LP13	CU13
정분류율	0.4996	0.9996	0.9824	0.9689	0.9502	0.9844	0.6626	0.7733	0.8354	0.7435	0.8354	0.8369
순위	6	1	3	4	5	2	6	5	2	4	2	1

LA__: LA 방법 이용, GW__: GW 방법 이용, LP__: Lepage 검정 이용, CU__: Cucconi 검정 이용,
K-S__: Kolbogorov-Smirnov 검정 이용, CVM: Craimer-von Mise 검정 이용

제 5 장 결론

본 논문에서는 새로운 픽셀값을 적절한 그룹으로 할당하는 이미지 픽셀 분류의 문제에서 가설 검정 기반의 LA 방법, 위치모수만을 고려한 GW 방법보다 위치모수와 척도모수의 동시검정법인 Lepage 검정, Cucconi 검정, K-S 검정, CVM검정방법이 이미지 픽셀의 정분류율이 높을 것으로 제안하였다.

모의실험을 통하여 위치모수와 척도모수의 동시검정법이 LA 방법이나 GW 방법보다 더 높은 정분류율을 가지는 것을 확인할 수 있었다. 주요 결과를 살펴보면 다음과 같다.

첫째, 총 두 가지의 모의실험 설계 중 첫 번째인 두 집단의 분산이 동일한 경우에는 위치모수만을 고려한 GW방법이 가장 우수하였으나 위치모수와 척도모수의 동시검정법 또한 높은 정분류율을 가졌다.

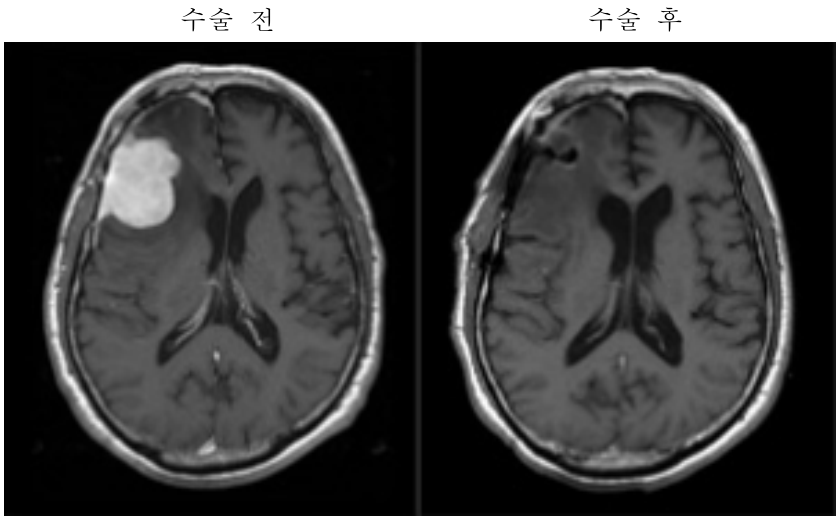
둘째, 두 번째 모의실험인 한 집단의 분산이 매우 커서 두 집단의 픽셀값이 겹치는 부분이 존재하는 경우, 위치모수와 척도모수의 동시검정법을 이용한 정분류율이 80%이상으로 높음을 확인하였다.

셋째, 분산이 같은 경우엔 GW방법의 정분류율이 높으며, 분산이 다른 경우에는 위치모수와 척도모수의 동시검정법 중 Cucconi 검정법을 이용한 이미지픽셀의 분류가 정분류율이 가장 높았다.

이미지 픽셀 분류는 의료분야, 얼굴인식 분야 등 다양한 분야에서 사용된다. 분류에 자주 사용되는 이미지 중 [그림 12][4]는 뇌종양 환자의 수술 전후 MRI 사진이다. 종양의 색과 정상인 부분의 색이 달라 픽셀값의 평균이 확연히 차이가 날 것으로 보인다. 이 경우, GW 방법과 위치모수와 척도모수의 동시검정법 모두 분류를 잘 할 것이라 예상할 수

4) http://www.gnmaeil.com/news/articleView.html?idxno=199630

있다.

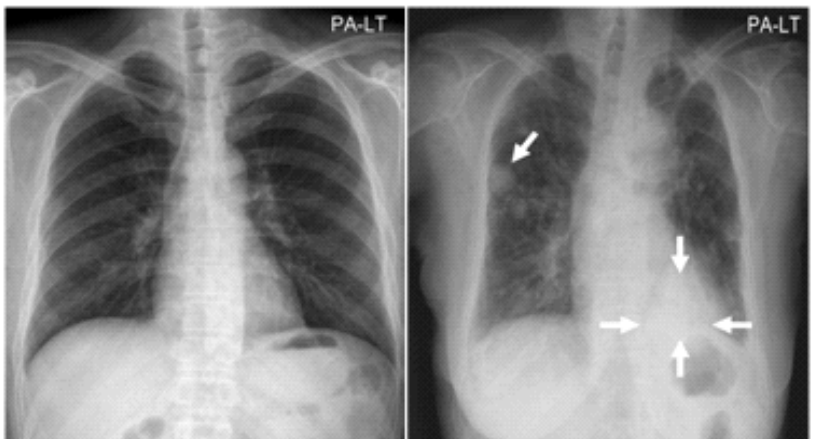


[그림 12] 뇌종양 환자의 수술 전후 MRI 사진

[그림 13][5]은 폐암환자와 정상인의 흉부 x-ray사진이다. x-ray 사진을 보면, 정상인에 비해 폐암환자의 흉부사진은 더 뿌옇게 흐려 보이며, 많은 경우 악성종양은 표면이 거칠어 그 주변의 이미지 픽셀값의 분산이 클 가능성 있다. 따라서 척추와 폐, 악성종양을 관심집단으로 설정한다면, 평균만을 고려했던 GW 방법 보다는 위치모수와 척도모수의 동시검정법을 이용하여 분류하는 것이 더 높은 정분류율을 보일 것이다.

5) 한림대학교 동탄 성심병원 호흡기센터:
http://dongtan.hallym.or.kr/RHC/?scr=411&mcode=01&scode=04&bno=1751

정상인 폐암환자



[그림 13] 폐암환자와 정상인의 흉부 x-ray사진

본 논문의 관심사는 분류하고자 하는 두 집단 중 분산이 큰 집단이 존재할 때, 새로운 이미지 픽셀값의 정분류율을 높일 수 있는지 여부였다. 논문에서 제안한 바와 같이 이미지 픽셀값의 분류방법으로 위치모수와 척도모수의 동시검정법을 이용한다면 정분류율을 높일 수 있다.

향후에는 분류에 관심 있는 집단이 세 집단 이상인 경우 이미지 픽셀의 정분류율을 높일 수 있는 방법에 대한 연구가 필요하다. 또한 색깔이 있는 3차원 이미지 픽셀값을 관심 있는 집단으로 분류하는 방법 연구의 필요성이 있다.

참고문헌

[1] 송문섭, 박창순, 김홍기, 2015. 비모수 통계학: R과 함께. 자유아카데미.

[2] Anderson, T.W., 1962. On the distribution of the two-sample Cramer-von Mises criterion. The Annals of Mathematical Statistics, 33(3), pp. 1148-1159.

[3] Ansari, A.R. & Bradley, R.A., 1960. Rank-sum tests for dispersions. The Annals of Mathematical Statistics, 31(4), pp. 1174-1189.

[4] Breiman, L. & Friedman, J. & Olshen, R. & Stone, C. J., 1984, Classification and regression trees. Belmont, CA: Wadsworth

[5] Cucconi, O., 1968. Un Nuovo Test Non Parametrico per il Confronto tra Due Gruppi Campionari, Giornale degli Economisti, XXVII, 225-248.

[6] Darling, D.A., 1957. The kolmogorov-smirnov, cramer-von mises tests. The Annals of Mathematical Statistics, 28(4), pp. 823-838.

[7] Ghimire, S. & Wang, H., 2012. Classification of image pixels based on minimum distance and hypothesis testing. Computational Statistics & Data Analysis, 56(7), pp. 2273-2287.

[8] Lee. A, J. & Hong A, C., 2013. Nonparametric Detection Methods against DDoS Attack. Korean Journal of Applied Statistics, 26(2), pp. 291-305.

[9] Lepage, Y., 1971. A combination of Wilcoxon's and Ansari-Bradley's statistics. Biometrika, 58(1), pp. 213-217.

[10] Liao, S. & Akritas, M., 2007. Test-based classification: A linkage between classification and statistical testing. Statistics & probability letters, 77(12), pp. 1269-1281.

[11] Marozzi, M., 2009. Some notes on the location - scale Cucconi test. Journal of Nonparametric Statistics, 21(5), pp. 629-647.

[12] Smirnov, N. V., 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull. Moscow Univ. 2, 3-16

[13] Vapnik, V. N., 1982. Estimation of Dependencies Based on Empirical Data. New York: Springer.

Abstract

A study on classification of image pixels based on statistical location-scale tests

Minjoo Lee
Department of Statistics
The Graduate School
Sungshin Women's University

Images can be considered as a finite group of pixel values. Thus images can be realized by groups of pixel values representing different regions in the image. The pixels representing a particular area or color in the images show more homogeneity regarding distribution of pixel values. Groups of similar image pixels can be classified by comparing pixels with each other and to pixels of known identity.

Liao & Akritas(2007) suggested a classification of image pixels based on hypothesis testings. This is a powerful nonparametric classification method. However, their method may misclassify many image pixels in the given image due to small p-values. So, Ghimire & Wang(2012) suggest new method for classification of image

pixels. But, their method also has a problem when variances of classes are different.

Thus, we have suggested a method for classification of image pixels based on location-scale tests(Kolmogorov-Smirnov test, Cramér-von Mises test, Lepage test, Cucconi test).

We simulated 6 methods in classifying of image pixels. When the population distributions have the same variances, classification rate of GW method is the highest. However, we discovered that classification rate of location-scale tests get excellent results when the population distributions haven't the same variances. Among them Cucconi test is the best classification rate.