확률적 접근을 사용한 피부 검출 방법

진영화

빅데이터 연구실

요약. 본 보고서에는 피부 검출을 위한 확률적 모형을 사용하고, 각각의 결과를 살펴보았다. 피부 검출 모형으로는 구별 모형 중 로지스틱 회귀모형과 히스토그램을 통해 추정된 확률질량함수를 사용한 ML 모형, 가우시안 분포를 이용한 ML 모형, 가우시안 분포와 conjugate prior 분포를 이용한 MAP 모형을 사용하였다. 결과는 정확도만을 고려하였을 경우 구별 모형의 성능이 좋았으나, 정밀도와 재현율을 고려한 전체적인 성능은 가우시안 기반의 생성모델이 더 우위에 있었다.

**Keywords: 컴퓨터 비전, 기계학습, 피부 검출, 구별 모형, 생성 모형**

1 서론

이미지 분류 문제는 이전부터 활발하게 연구되어 오던 컴퓨터 비전 분야의 문제 중 하나이다.

컴퓨터 비전에서 이미지 분류를 하기 위한 접근법은 크게 두 가지로 나눌 수 있다. 하나는 데이터가 주어졌을 때 데이터에 대한 정답이 나타날 조건부 확률을 직접적으로 반환하는 모형이다. 이를 구별 모형 (Discriminative Model)이라고 하며, 로지스틱 선형 회귀, 서포트 벡터 머신, 딥 러닝 등 많이 알려진 기계학습 모형 등이 이에 속한다. 구별 모형은 가정이 비교적 단순하며, 데이터의 양을 충분히 확보할 수 있다면 좋은 성능을 기대할 수 있다. 이와 달리, 생성 모형(Generative Model) 이라고 불리는 접근 방법은 데이터들의 분포와, 그 분포로부터 주어진 데이터가 생성되었을 확률을 이용하여 데이터를 분류하는 모형이다. 생성 모형은 구별 모형에 비해 데이터의 분포에 대한 가정들이 추가로 필요한데, 이 가정이 잘못되었을 경우 구별 모형에 비해 성능이 떨어질 수 있다. 하지만 가정이 잘 구축되어 있다면, 이상치에 강건하고 학습 데이터가 비교적 적은 상황이라도 좋은 예측 성능을 보일 수 있다.

본 보고서에서는 여러 이미지 분류 문제 중, 이미지상의 픽셀 중 피부인 픽셀을 검출해내는 문제를 다양한 확률 모형을 사용하여 결과를 살펴보고자 한다. 데이터셋은 Face Detection Data Set 2,845장을 사용하였다.

본 보고서는 다음과 같이 진행된다. 1장에서는 연구 목적과 본 보고서의 구성에 대해 간략히 소개하고, 2장에서는 피부 검출을 위해 사용할 로지스틱 회귀모형, 최대 가능도 추정모형, 최대 사후확률 추정모형에 대해 설명한다. 3장에서는 각 모형을 사용한 피부 검출 결과를 상세하게 설명한다. 4장에서는 실험결과를 토대로 결론을 도출한다.

2 확률 추론 모형

2.1 로지스틱 회귀 모형

로지스틱 회귀 모형은 구별 모형(Discriminative Model)에 속하는 확률 모형이다. 구별 모형(Discriminative Model)이란, 데이터 가 주어졌을 때 해당 데이터가 정답 클래스 에 속할 확률 을 직접 계산하는 모형이다. 구별 모형은 데이터에 대한 가정이 비교적 단순하여, 데이터를 많이 확보할 수 있다면 일반적으로 좋은 성능을 기대할 수 있다.

이산형 변수 예측 문제에서, 가장 쉽게 사용할 수 있는 구별 모형 중 하나인 로지스틱 회귀 모형은 기존의 회귀분석이 이산형 변수를 예측하는 문제에 사용하기에는 적합하지 않기 때문에 기존 회귀분석의 종속변수를 베르누이 시행으로 모형화한 회귀분석 방법이다.

선형회귀는 독립변수 와 연속형 종속변수 의 관계를 선형으로 가정하고 이를 가장 잘 표현할 수 있는 회귀계수를 데이터로부터 추정하는 모형이다. 설명 변수가 개인 다중선형회귀의 일반 식은 다음과 같다.

|  |  |
| --- | --- |
|  | (**1**) |

하지만 종속변수 가 0과 1로 나타나는 이진변수 일 경우, 선형회귀의 추정 값 가 실제 종속변수의 범위와 일치하지 않기 때문에, 선형회귀모형을 곧바로 적용할 경우 성능을 기대할 수 없다. 따라서 이를 해결하기 위해서는 를 의 값을 가지는 베르누이 확률변수로 가정하여 모형을 새롭게 재정의할 필요가 있다. 하지만 곧바로 종속변수 를 이진 확률 변수로 가정할 경우 식 (**1**)에서 양변의 출력 값이 일치하지 않는 문제가 발생한다. 따라서 이를 해결하기 위해, 승산(odds)이라는 개념을 사용한다. 승산(odds)은 임의의 사건이 발생하지 않을 확률 대비 발생할 확률의 비율을 뜻하는 개념으로, 다음과 같이 정의한다.

|  |  |
| --- | --- |
|  | (**2**) |

승산은 의 범위를 가지는데, 이를 실수 전체로 확장하기 위해 로그를 취한다. 이를 로짓(logit)이라고 정의하며, 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | |  | (**1**) | | (**3**) |

로짓은 의 범위를 가지기 때문에, 식 (**1**)에서의 우변과 범위가 일치하므로, 선형회귀를 다음과 같이 모형화 할 수 있다.

|  |  |
| --- | --- |
|  | (**4**) |

식 (**4**)를 에 관한 함수로 정리해주면, 식은 다음과 같아진다.

|  |  |
| --- | --- |
|  | (**5**) |

식 (**5**)에서 우변에 도출된 식을 시그모이드(sigmoid) 혹은 로지스틱(logistic) 함수라고 부르며, 시그모이드 함수를 이용하여 이진 종속변수의 값을 예측하는 확률모형을 로지스틱 회귀 분석이라고 정의한다. 로지스틱 회귀분석에서 모수는 벡터 이며, 주어진 데이터를 통해 가장 적합한 를 학습하는 방향으로 모형을 만든다. 이 보고서에서는 다음 생성 모형(Generative Model)과의 결과 비교를 위한 베이스라인으로써 로지스틱 회귀 모형을 사용하였다.

2.2 최대 가능도 추정 모형

클래스 구별 문제에서, 관측된 데이터 가 클래스 에 속할 확률을 라고 할 때, 베이즈 정리에 의해 클래스 에 속할 확률을 다음과 같이 표현할 수 있다.

|  |  |
| --- | --- |
|  | (**6**) |

식 (**6**)에서 우변의 분모 는 식에 영향을 주지 않는 상수이므로, 결국 주어진 데이터가 특정 클래스에 속할 확률 는 클래스 자체의 확률 와 데이터가 해당 클래스에서 생성될 확률 의 곱으로 표현됨을 의미한다. 이처럼 를 직접 구하지 않고, 두 확률 분포 와 를 통해 간접적으로 계산하는 모형을 우리는 생성 모형(Generative Model)이라고 부른다. 이 때 는 사전 확률(prior), 는 가능도(likelihood)를 의미하며, 이 둘의 결합확률분포를 사후 확률(posterior)라고 부른다.

는 일반적으로 특정 분포를 가정하게 되는데, 이 때 모수는 최대 가능도 추정(Maximum Likelihood Estimation) 방법을 사용하여 찾게 된다. 최대 가능도 추정법은 알 수 없는 모수 값 를 가지는 확률 분포에서 뽑은 데이터 들을 바탕으로 가장 확률이 높은 값 를 추정하는 방법이다. 여기서 가능도(likelihood)의 사전적 정의는 이미 주어진 데이터 들에 비추어 보았을 때 추정한 모수 가 실제 모수 값 일 정도를 가리키는 값이다. 만약 확률 변수 가 로 주어져 있고, 확률 변수 가 확률 분포로 를 가진다면, 가능도함수는 다음과 같이 표현할 수 있다.

|  |  |
| --- | --- |
|  | (**7**) |

가능도함수는 곱셈연산으로 이루어져 있어 계산의 과부하가 심하기 때문에, 계산의 편의성의 위해 로그를 취한 값을 많이 사용하며, 이를 로그가능도라 정의하고 다음과 같이 표현할 수 있다.

|  |  |
| --- | --- |
|  | (**8**) |

최대 가능도 추정법은 위의 가능도함수(또는 로그가능도함수)를 다음 식 (**9**)와 같이 최대화하는 값 을 우리가 찾는 모수 로 삼는 방법이다.

|  |  |  |
| --- | --- | --- |
|  | (**9**) |  |

최대 가능도 추정 모형에서는 먼저 가정한 분포 하에서의 가능도 함수를 정의한 후 위의 최대 가능도 추정법을 통해 를 찾는다. 이후 를 통해 계산한 이용하여 주어진 데이터가 클래스에 속할 확률을 구하게 된다. 이후 이러한 과정을 모든 클래스에 걸쳐 계산하고, 이 중 가장 높은 확률이 나온 클래스로 데이터 를 분류하게 된다.

2.3 최대 사후 확률 추정 모형

식 (**6**)에서 를 모수 를 따르는 특정 분포로 가정할 경우 다음과 같이 표현할 수 있다.

|  |  |
| --- | --- |
|  | (**10**) |

이 때 모수 에 대한 확률 는 일반적으로 균일하게 가정한다. 이는 데이터에서 어떤 가 추정될 지 알 수 없기 때문이다. 하지만 이 를 다시 특정 확률분포에서 얻을 수 있는 확률변수로 가정한 뒤 사후확률을 구할 수 있는데, 이런 방법을 최대 사후 확률 추정(Maximum A Posteriori)이라고 부른다.

가능도 함수의 모수 에 대한 사전확률 를 특정분포로 가정할 때는 일반적으로 가정한 가능도함수와 Conjugate Relationship을 가지는 분포를 선택하는데, 이는 결합되는 두 분포가 Conjugate Relationship을 가질 경우 이미 알려진 다른 특정 분포의 형태로 표현될 수 있기 때문이다. 예를 들어 데이터의 분포 를 다변수 가우시안 분포로 가정할 경우, 일반적으로 의 분포 는 가우시안 분포와 Conjugate Relationship을 가지고 있는 정규-역-위샤트 분포(Normal Inverse Wishart, NIW)를 가정하게 된다. 이는 가우시안 분포의 모수 ()를 NIW분포의 확률변수로 가정한다는 말과 같다.

|  |  |
| --- | --- |
|  | (**11**) |

이 두 분포를 결합하게 되면, 다음과 같은 t 분포로 다시 표현할 수 있다.

|  |  |
| --- | --- |
|  | (**12**) |
|  |  |

이렇게 식 (**12**)에 볼 수 있듯이, 두 분포가 Conjugate Relationship을 가지고 있다면 또 다른 분포의 형태로 표현할 수 있게 된다. 만약 서로 Conjugate Relationship관계가 없는 분포라면 이러한 닫힌 형태(Closed-Form)로 나타나지 않기 때문에, 수치해석적인 방법을 통해 근사계산을 하여야 한다.

이상 설명한 3가지 모형을 기반으로 본 보고서에서는 실험을 진행하였다.

3 실험 결과

3.1 실험 환경

피부 검출을 위한 데이터셋은 Face Detection Data Set 2,845장을 사용하였다. 데이터 이미지들의 각각의 픽셀 RGB 값을 독립변수 로 보고, 이에 대한 피부 여부 를 종속 변수로 설정하였다. 계산비용 상의 문제로, 전체 이미지중 1000장의 이미지를 임의선택한 후 다시 약 700장 : 300장으로 나누어 각각 트레이닝셋과 테스트셋으로 분류하여 실험하였다. 그 후 이러한 과정을 로지스틱 회귀는 100번, 생성모형은1000번 반복한 후 평균 정확도를 측정하는 교차 검증(cross validation)을 사용하였다. 학습은 R 채널만을 이용한 경우와 RGB 채널을 이용한 경우로 나누어 진행하였다. 모델들의 성능은 정확도와 정밀도(Precision), 재현율(Recall)로 측정 ∙ 비교하였다

실험은 먼저 데이터를 분포를 간단히 살펴본 뒤, 로지스틱 회귀 모형, 히스토그램을 통해 추정한 확률질량함수를 사용한 생성 모형, 가우시안 분포를 사용한 생성 모형, 가우시안과 사전분포의 결합분포를 사용한 MAP 생성모형의 순서대로 살펴본다.

3.2 데이터 분석

모형 제작에 앞서 피부인 픽셀(스킨 픽셀)과 그렇지 않은 픽셀(논스킨 픽셀)에 대한 분석을 수행하였다.

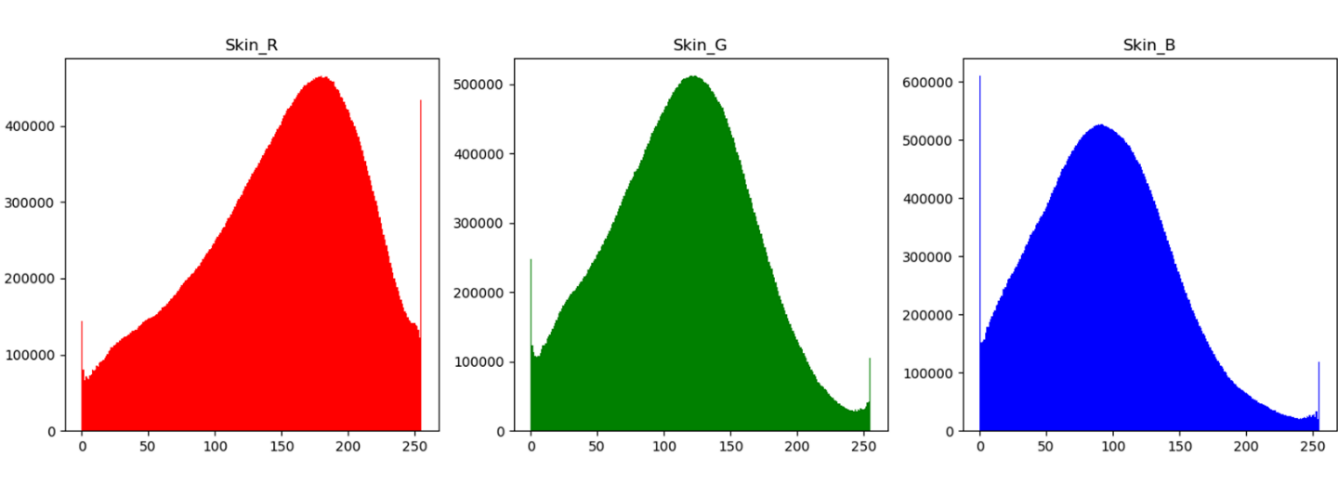
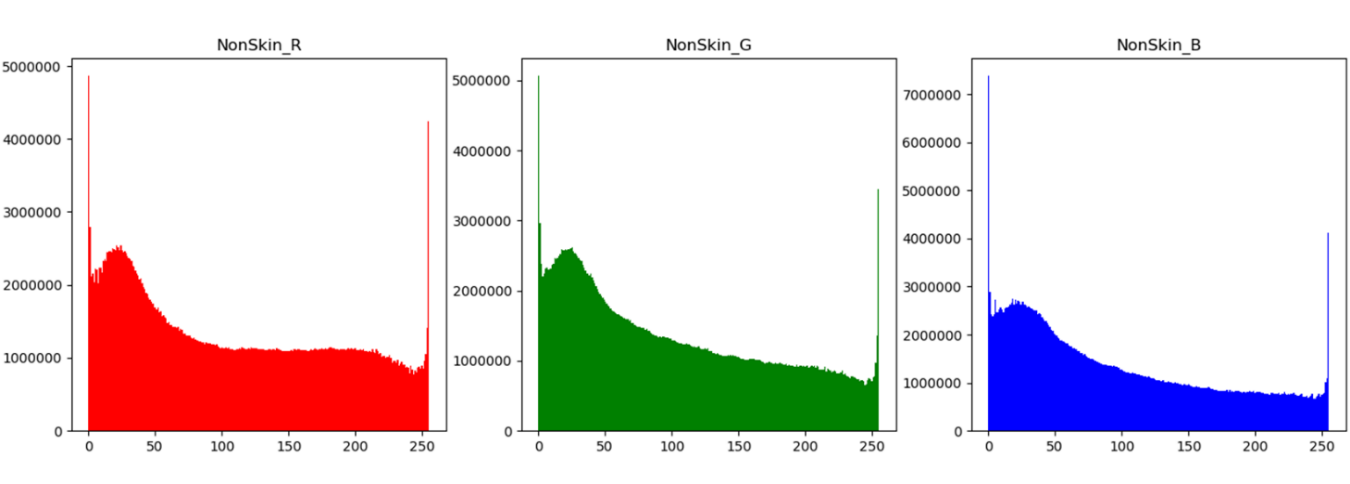


그림 1. 전체 데이터의 피부인 픽셀과 피부가 아닌 픽셀들의 분포

그림1. 을 살펴보면, 피부인 픽셀과 피부가 아닌 픽셀의 분포에 차이가 있다는 점을 알 수 있다. 두 분포 간의 각 채널별 최빈값의 차이가 R→G→B 순서로 많이 나고 있다. 그리고 모든 채널에서 0과 255값이 다른 값들에 비해 많이 나타나고 있다는 점 또한 확인할 수 있다.

3.3 실험 결과

로지스틱 회귀모형을 이용한 실험에서는 계산 비용상의 문제로 RGB값만을 사용하여 실험하였다. 생성모형을 이용한 실험에서는 에 대한 분포에 대해, 관측된 데이터의 히스토그램을 통해 추정한 확률질량함수(Probability Mass Function), 가우시안 분포(Gaussian distribution), 가우시안과 사전 분포를 결합한 결합 분포를 이용해 각각 가정한 후 실험을 진행하였다. 확률질량함수는 각 채널에서 0~255 값들 각각의 비율을 확률로 사용하였다.

|  |  |  |
| --- | --- | --- |
|  | (**13**) |  |

이후 각 채널을 독립으로 가정한 후, 각 채널에서의 확률의 곱을 통해 피부일 확률과 그렇지 않을 확률을 계산하였다.

가우시안 분포에서는 분포의 MLE 값인 피부인 픽셀들과 피부가 아닌 픽셀들 각각의 ( 벡터)와 ( 매트릭스)를 모수로 사용하였다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | R | | | RGB | | |
| 모형 | Acc | Pre | Rec | Acc | Pre | Rec |
| 로지스틱 | . | . | . | 0.81 | 0.18 | 0.02 |
| 확률질량 | 0.58 | 0.25 | 0.77 | 0.62 | 0.27 | 0.74 |
| 가우시안 | 0.57 | 0.24 | 0.76 | 0.76 | 0.38 | 0.82 |
| 가우시안 MAP | 0.57 | 0.24 | 0.76 | 0.76 | 0.38 | 0.82 |

가우시안 분포와 사전 분포를 이용한 MAP 방법에서는, R 채널만을 이용했을 때는 가우시안과 정규-역-감마분포(Nomal Inverse Gamma, NIG)를, RGB채널을 이용했을 때는 가우시안과 정규-역-위샤트분포(Normal Inverse Wishart, NIW)를 사용하였다. 이 때 NIG의 파라미터는 을, NIW의 파라미터는 을 사용하였다. 결과는 표1에 정리하였다.

표 1. 확률 모형을 사용한 피부 검출 결과

표 1의 결과를 보면, 로지스틱 회귀모형의 경우, 정확도는 사용된 모형중 가장 높았지만, Precision과 Recall의 비율이 다른 모델에 비해 많이 떨어지는 것을 확인할 수 있다. 이는 데이터 불균형으로 인한 문제로, 학습 이미지에서 피부인 픽셀보다 그렇지 않은 픽셀의 비율이 훨씬 많기 때문에 모델 자체가 데이터를 논스킨으로 구별할 확률이 높아져 버렸기 때문이다. 이로 인해 정확도 자체는 높지만 스킨을 검출하는 능력은 확연히 떨어져 버린 것으로 보인다. 이를 통해 구별모델은 데이터의 비율도 같이 학습한다는 특징을 확인할 수 있다.

생성모형의 결과를 살펴보면, R 채널만 이용하였을 때는 3가지 모형의 성능이 비슷하였다. 이는 R 채널에서의 픽셀의 분포가 가우시안 분포에 가깝게 나타나기 때문이라고 생각해 볼 수 있다. RGB 채널을 모두 이용하였을 경우에는 가우시안 기반의 두 모델의 성능이 확률질량함수를 사용한 모델보다 결과가 좋은 것을 알 수 있다. 이는 확률질량함수의 경우 각 채널을 독립으로 가정하였으나, 가우시안분포의 경우 각 채널의 공분산을 고려하였기 때문으로 생각해 볼 수 있다. 마지막으로 가우시안과 가우시안 MAP 방법이 성능이 똑같이 나오는 것을 확인 할 수 있는데, 이는 굉장히 많은 데이터 인스턴스들로 인해 사용자가 부여한 사전지식이 거의 반영되지 않았기 때문으로 볼 수 있다.

3.4 사전확률을 조정한 실험 결과

생성모델의 장점 중 하나는 사전확률을 통해 사용자의 지식을 모형에 반영할 수 있다는 점이다. 이 점을 이용하여 각각의 클래스에 대한 사전확률을 다르게 준 후 다시 실험을 진행하였다. 이 때 픽셀이 피부일 사전 확률은 로 가정하였다. 결과는 표 2에 정리하였다. 사전확률을 조절한 결과, 전반적으로 정확도와 Precision이 상승하고. Recall의 성능이 낮아진 것을 볼 수 있다. 이는 이미지 픽셀을 스킨으로 예측할 확률이 낮아짐으로써, False Positive의 비율은 낮아졌지만, 이와 함께 스킨을 검출해내는 능력 또한 낮아짐으로써 False Negative가 늘어난 결과로 볼 수 있다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | R | | | RGB | | |
| 모형 | Acc | Pre | Rec | Acc | Pre | Rec |
| 확률질량 | 0.68 | 0.27 | 0.56 | 0.66 | 0.29 | 0.69 |
| 가우시안 | 0.66 | 0.25 | 0.54 | 0.81 | 0.46 | 0.75 |

표 2. 임의의 사전확률을 준 모형의 피부 검출 결과

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | R | | | RGB | | |
| 모형 | Acc | Pre | Rec | Acc | Pre | Rec |
| 확률질량 | 0.56 | 0.24 | 0.77 | 0.59 | 0.25 | 0.75 |
| 가우시안 | 0.58 | 0.24 | 0.75 | 0.76 | 0.41 | 0.81 |

표 3. 극단값이 제거된 데이터를 학습한 확률 모형의 피부 검출 결과

앞서 살펴보았던 데이터 분포에서, 모든 채널에서 {0, 255} 두 값이 다른 값들에 비해 상당히 많이 관측되었던 것을 확인할 수 있었다. 극단값에 영향을 많이 받는 평균과 분산의 특성상 이 점이 영향을 미칠 수 있다고 판단하여, 학습 데이터에서 0과 255이 값을 제거한 후 다시 모형을 학습시켜 실험을 진행하였다. 결과값은 표 3에 정리하였다.

표 3에서 살펴볼 수 있듯이, 전반적인 성능상의 차이에는 큰 변화가 없었다. 오히려 확률 질량 함수는 정확도가 소폭 하락하는 모습 또한 확인할 수 있었다. 이는 상당히 많은 학습 데이터로 인해 {0, 255} 값이 실제 생성모형의 모수를 추정하는데 크게 영향을 미치지 않는다는 점을 말해준다.

4 결론

본 보고서에서는 컴퓨터 비전의 여러 문제 중 하나인 피부 검출 문제를 해결하는 과정을 정리하였다.

구별모델 중 하나로 사용하였던 로지스틱 회귀 모형은 많은 가정없이 사용할 수 있는 점이 장점이었지만, 데이터에서 두 클래스의 비율차이가 현저하게 나기 때문에 정밀도와 재현율 측면에서 좋지 않았다.

생성모델은 정밀도와 재현율의 측면에서 구별모델보다 성능적으로 우위를 보였다. R 채널만을 사용하였을 때는 확률 질량 모형과 가우시안 모형의 성능이 유사했으며, 이 점을 통해 픽셀들의 분포가 가우시안 분포를 따르고 있다고 생각할 수 있다. 반면 RGB채널을 모두 이용하였을 때는 가우시안 모형의 성능이 더 우수했으며, 이는 RGB 채널들간의 상관관계를 고려하였기 때문으로 여겨진다. 한편 가우시안 분포와 conjugate prior분포를 이용한 MAP 모형의 경우, 사전 분포의 파라미터에 크게 영향을 받지 않는 모습을 보였다. 이는 학습 데이터가 상당히 많기 때문에 사전 확률의 지식이 거의 반영되지 않은 것이라고 볼 수 있다.

이후 성능향상을 위해, 모델에 사전확률을 조정한 경우와 데이터에 극단값을 제거한 후 학습을 진행한 경우로 나누어 다시 실험하였다. 사전확률을 조정한 경우, 정확도와 정밀도면에서는 상승하는 모습을 보였지만, 재현율 측면에서는 성능이 하락해 일장일단이 있었다. 반면 각 채널별로 {0, 255} 값을 제거하고 학습일 진행하였을 때는 큰 변화가 없었다. Lena 이미지를 이용한 모델 테스트 결과는 그림 2. 에서 확인할 수 있다.



그림 2. Lena 이미지를 사용한 피부 검출 테스트. 가장 상단의 이미지가 원본 이미지이며, 첫째 줄은 왼쪽부터 Pmf 모델, 가우시안 모델, 가우시안 MAP 모델, 둘째 줄은 사전확률을 조정한 Pmf 모델, 가우시안 모델, 셋째 줄은 {0, 255} 픽셀은 제거한 후 학습한 Pmf 모델, 가우시안 모델