선형결합을 이용한 패턴분류의 한 방법 - 이론

°김 경 태
한국기계연구소

A method of pattern classification using linear combination

Kim, Kyung Tae
Korea Institute of Machinery and Metals

1. 서론

패턴인식의 이론이 정립된 후 실제 공학에 응용되기 시작한 것은 전자계산기가 실용화되기 시작한 1970년대 초반부터이다. 그동안 전자계산기의 많은 발전으로 인해 실행속도가 많이 빨라졌지만 알고리즘과 이론분야에 있어서 실시간처리와 높은 인식율을 전제로 했을때 계산량을 생각지 않을 수 없다.

그래서 본 논문에서 계산량을 감소시키기 위한 한 방법을 제안하다.

일반적으로 패턴인식의 문제에 있어서 입력벡타의 차원이 높으면 높을수록 대부분의 경우 높은 인식율이 얻어진다. 그러나 그만큼 복잡한 계산과 상당한 계산량이 필요하다.

따라서 입력벡타를 두 부분으로 나누어서 낮은 차원인 두 종류의 변수를 도출하여 인식오차가 최소가 되게하기 위한 선형결합을 행한다.

이때 필요한 것이 선형결합을 위한 계수가 문제인데 본 논문에서는 이 계수를 구하는 방법에 대하여 논한다. 물론 이들 계수(혹은 하중함수)를 구하기 위해 이미 알고 있는데이타로써 학습(learning) 에 의한 방법을 이용할 수도 있었으나[1)-2)] 이 방법은 만약 선형분리가 안되는 샘플들의 분포에 대해서는 계수가 수렴하지 않고 발진하는 경향이 있기 때문에 사용하기 곤란하다.

또한 문제를 간단히 하기 위해 서로다른 두종류의 정보로써 두 카테고리를 분류하는 경우는 생각하기로 한다.

2. 본론

입력벡터 n 차원인 x 를 통계학적 패턴분류를 하고자 할 경우 n 이 클때는 그것들이 의미를 가지기 위해서는 그 만큼 많은 샘플수가 필요하다. 따라서n차원의 x를 d차원의 x_1 와 n-d 차원인 x_2 로 나누고 다변수 x_1, x_2 로부터 각각의 변수를 도출한 것을x_A, x_B로 생각한다. 그리고 카테고리의 수를 2로 하면 카테고리1의 A정보에 의한 확률변수를 x_{A1}, B정보에 의한 확률변수를 x_{B1} 로 하고, 카테고리 2에 대하여는 각각 x_{A2}, x_{B2}라 정의한다. 그리고 이들 확률변수들이 각각 독립이고 정규분포를 한다고 하면 카테고리1, 2의 확률변수는 다음과 같이 선형결합식으로 이루어진다.

$$\alpha x_{A1} + \beta x_{B2} = x_1 \quad (1)$$

$$\alpha x_{A2} + \beta x_{B2} = x_2 \quad (2)$$

그리고 각각 해당평균치와 분산으로 이루어진 확률밀도함수를 그림 1과 같이 나타내면

$$m_1 = \alpha m_{A1} + \beta m_{B1} \quad (3)$$

$$m_2 = \alpha m_{A2} + \beta m_{B2} \quad (4)$$

$$s_* = \alpha^2 s^2 + \beta^2 s^2 \quad (5)$$

로 나타낼수 있다. 여기서 분산은 같다고 가정한다.

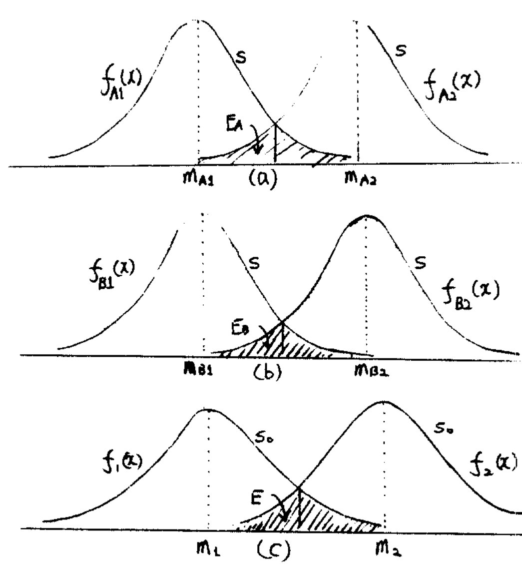

그림1: 각기 다른 확률변수에 의한 확률밀도함수

그림 1에서 $f_{A1}(x)$ 는 x_{A1} 에 대한 확률밀도함수를 나타내며 빗금친 부분은 오인식되는 부분이다.

x_A 를 이용했을때 오인식율 $\frac{1}{2}$ 을 E_A 그림1의 ((a)), x_B 에 대한 오인식의 $\frac{1}{2}$ 을 $E_B((b))$, 라 할때 x_A 와 x_B 의 선형결합으로 이루어져 형성된 분포중 오인식의 $\frac{1}{2}$ 을 $E((c))$ 라 하면 문제는 E 가 E_A 보다 적고, 또 E_B 보다 적도록 선형결합을 시키기 위한 α, β 를 구하는 문제에 귀결된다. 그러면

$$E_A = \int_{-\infty}^{\frac{m_{A2}-m_{A1}}{2}} \frac{1}{\sqrt{2\pi}\,S} e^{-\frac{(x-m_{A2})^2}{2S^2}} dx \qquad (6)$$

가 되며 분산이 1이고 평균치가 0가 되도록 정규화 하면

$$E_A = \int_{-\infty}^{P_A} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \qquad (7)$$

이고 E_B, E 도 마찬가지로

$$E_B = \int_{-\infty}^{P_B} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \qquad (8)$$

$$E = \int_{-\infty}^{P} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \qquad (9)$$

여기서,

$$P_A = \frac{m_{A1} - m_{A2}}{2S} \qquad m_{A2} > m_{A1} \qquad (10)$$

$$P_B = \frac{m_{B1} - m_{B2}}{2S} \qquad m_{B2} > m_{B1} \qquad (11)$$

$$P = \frac{m_1 - m_2}{2S_0} \qquad m_2 > m_1 \qquad (12)$$

도 된다.

그래서 $E < E_A$ 와 $E < E_B$ 가 되기 위해서는 $P < P_A$ 그리고 $P < P_B$ 가 되어야 하므로 식(3), (4), (5)를 이용하여 (10), (11), (12)를 풀면

$$\sin(\theta + \phi) > \frac{C}{\sqrt{C^2 + D^2}} \qquad (13)$$

$$\sin(\theta + \phi) > \frac{D}{\sqrt{C^2 + D^2}} \qquad (14)$$

여기서

$$C = m_{A2} - m_{A1} \qquad (15)$$

$$D = m_{B2} - m_{B1} \qquad (16)$$

$$\theta = \tan^{-1}\frac{\alpha}{\beta} \qquad (17)$$

$$\phi = \tan^{-1}\frac{D}{C} \qquad (18)$$

이다. 식(13), (14) 가 만족되기 위해서는 $\sin(\theta + \phi) = 1$ 이면 된다. 즉,

$$\theta + \phi = \frac{\pi}{2} \qquad (19)$$

이면 된다.

그러므로

$$\theta = \frac{\pi}{2} - \tan^{-1}\left(\frac{m_{B2} - m_{B1}}{m_{A2} - m_{A1}}\right) \qquad (20)$$

식(17)로부터

$$\tan\theta = \frac{\alpha}{\beta} \qquad (17)'$$

이므로 선형결합계수 α, β 의 관계가 비율로서 얻어진다.

3. 결론

본 논문에서 고차원의 다변량을 그대로 이용하기에는 많은 시간과 계산량이 필요하다. 이들은 두 종류의 변량으로 나누어서, 나누어진 수변량을 다시 묶기 위해 선형결합을 시도했다. 이때 선형결합을 위한 계수 α, β 를 구하는 문제를 다루었는데 기본적인 사항에 대해 알아보기 위해 두 종류의 변수와 카테고리별로 분산이 같다는 사실을 전제로 했다.

결과적으로 식(17)'과 같이 α, β 사이의 관계는 $\tan\theta$ 에 의한다는 사실을 발견했다.

물론 $\tan\theta$ 를 구하기 위해서는 이미 알고 있는 샘플들의 평균치을 계산한다. 그러면 실제로 미지의 패턴 x 에 대해 x_A 와 x_B로 분리시켜 선형결합을 한후 그것을 카테고리 1의 확률밀도함수와 카테고리 2의 확률밀도함수에 적용시켜 x 가 속하는 카테고리를 결정하면 된다.

마지막으로 본 논문에서는 각 샘플분포의 분산이 같다는 사실을 전제로 했으나 이 사실은 실제로 극히 드문 일이라 생각되며 앞으로 분산이 다른 경우에 대해서도 생각해야 할 것이다.

감사의 글: 평소의 연구에 많은 격려를 하시고 한국경제연구소 이해소장님과 연세대학교 전자공학과 차일환 교수님께 감사를 드립니다.

References

1. R.O Duda & P.E. Hart, Pattern Classification and scene analysis, chapter 1-5, John Wiley & sons, 1973
2. N.J. Nilson, Learning Machines, Mc Graw-Hill Book, 1965