4. 분류 기법: 마할라노비스-다구찌 시스템

4.1 개요

분류 알고리즘(classification algorithm) 이란 주어진 데이터를 특정 목적에 맞게 분류하는 알고리즘을 의미한다. 예를 들어서, 개와 고양이를 분류하는 알고리즘, 지문이나 사진으로 사용자를 판단하는 알고리즘, 공장 설비가 정상으로 작동하는지를 감시하는 알고리즘 등이 있으며, 최근 딥러닝(deep learning)의 등장으로 엄청난 성과를 내고 있는 분야이다.

다음 그림은 대표적인 분류 알고리즘의 하나인 결정 트리에 쓰이는 훈련 데이터의 일부이다.

입력				출력
Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	79	90	Strong	No
Sunny	56	70	Weak	Yes
Sunny	79	75	Strong	Yes
Sunny	60	90	Strong	No
Overcast	88	88	Weak	No
Overcast	63	75	Strong	Yes
Overcast	88	95	Weak	Yes
Rain	78	60	Weak	Yes
Rain	66	70	Weak	No
Rain	68	60	Strong	No

그림 4.1 분류를 위한 훈련 데이터의 일부

이 데이터의 목적은 모델을 만들어 새로운 데이터를 넣었을 때 테니스를 칠 수 있는지 여부를 묻는 것이다. 이때 입력은 Outlook(날씨), Temperature, Humidity, wind가 되고 출력은 Play Tennis의 Yes, No이다.

여기서 Play Tennis의 Yes나 No와 같은 것을 레이블(lable)이나 클래스(class)라고 한다.

이 데이터에는 Outlook(날씨), Temperature(온도), Humidity(습도), Wind(바람), Play Tennis(테니스 하기) 와 같이 5개의 속성(attribute)이 있다. 그리고 Outlook이 가지고 있는 데이터를 보면 Sunny, Overcast, Rain과 같이 숫자가 아닌 데이터를 갖고 있다.

이와 같은 데이털르 **범주형 자료(categorical data)**라고 하며, Wind와 Play Tennis도 같은 범주형 자료를 가지고 있다. 반면에, Temperature, Wind는 익숙한 숫자 데이터를 가지고 있다. 이것을 **수치형 자료(numerical data)**라고 한다.

분류 알고리즘은 다음과 같이 훈련할 수 있다.

먼저, 알고리즘에 분류하고 싶은 훈련 데이터(Training data)를 주고 훈련한 후에 테스트 데이터(Test data)를 주어 훈련 결과를 확인하는 것이다.

이때 분류 알고리즘의 학습은 두 가지 종류가 있다. Play Tennis 와 같이 클래스가 Yes, No 두 가지 중에 하나를 학습해야 하는 것을 이진 분류(binary classification)라고 하며, 개, 고양이, 토끼 처럼 여러 클래스를 학습해야 하는 것을 다중 클래스 분류(multi-class clssification)라고 한다.

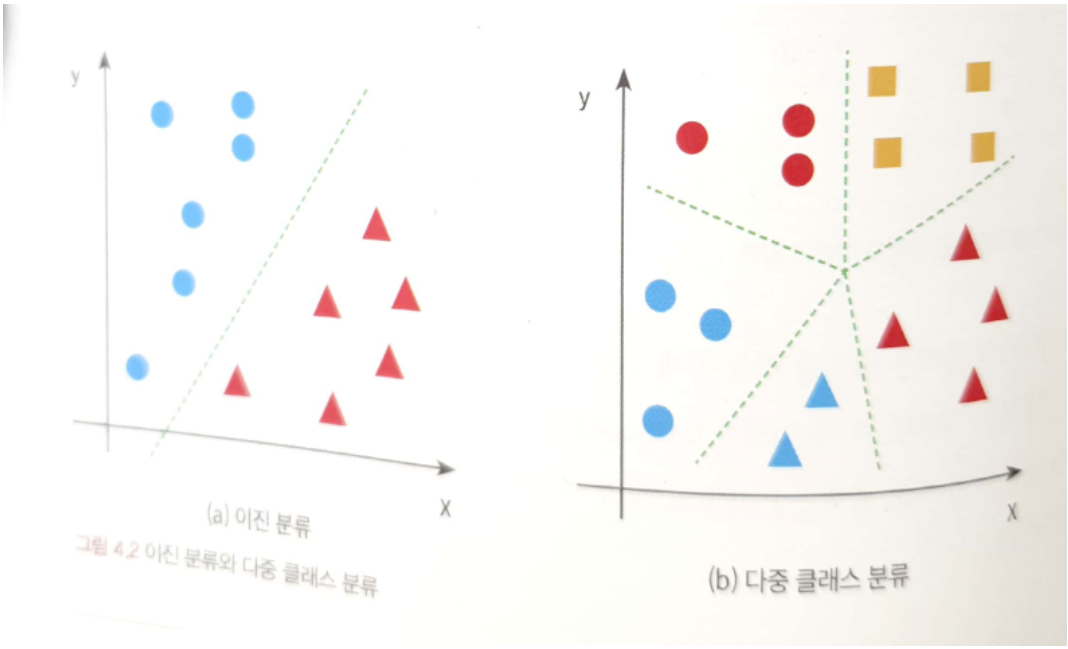


그림 4.2 이진 분류와 다중 클래스 분류

그림 4.2는 이진 분류와, 5개를 분류하는 다중 클래스 분류의 이미지를 보여준다.

다중 클래스 분류는 신경망, 결정 트리와 같이 한 번에 다중 클래스를 학습시켜서 분류를 하는 방법과 여러 개의 이진 클래스로 나눠서 학습을 시키는 방법이 있다. 이때 각 분류 알고리즘의 특징을 파악해 사용하는 것이 중요하다.

이진 분류는 가장 기초적이며 실용적인 알고리즘이다. 정상/비정상, 좋음/나쁨, 비/갬 등 세상의 모든 문제는 이진 분류 문제로 표현 가능하다.

이런, 이진 분류 문제도 두 가지로 분류할 수 있다.

개와 거북이, 남자와 여자 등 명확히 나눌 수 있는 두 개의 클래스를 분류하는 것과 정상/비정상과 같이 클래스는 두 개이지만 명확히 나눌 수 없는 두 개의 클래스를 분류하는 것이다.

여기서 정상/비정상이 명확히 나눌 수 없는 두 개의 클래스라는 이유는 다음과 같다.

개와 거북이, 남자와 여자는 두 가지 다 정확하게 정의할 수 있다. 하지만 정상 상태와 비정상 상태는 매우 비슷한 것이 특징이다.

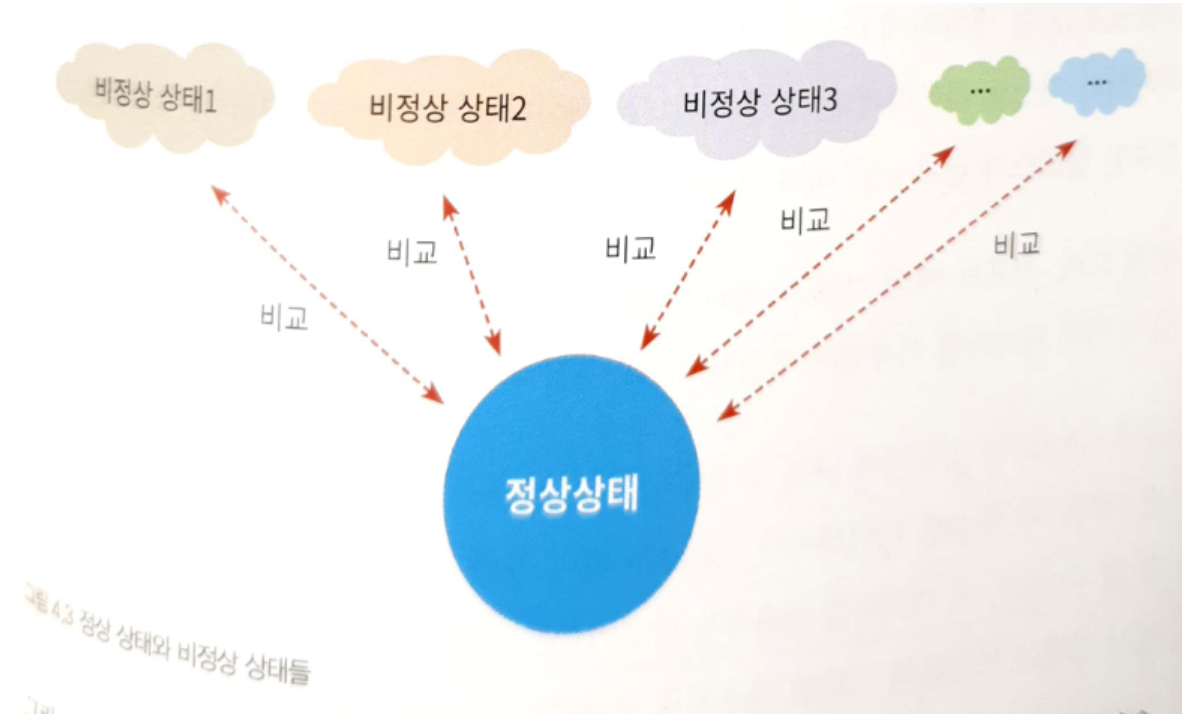
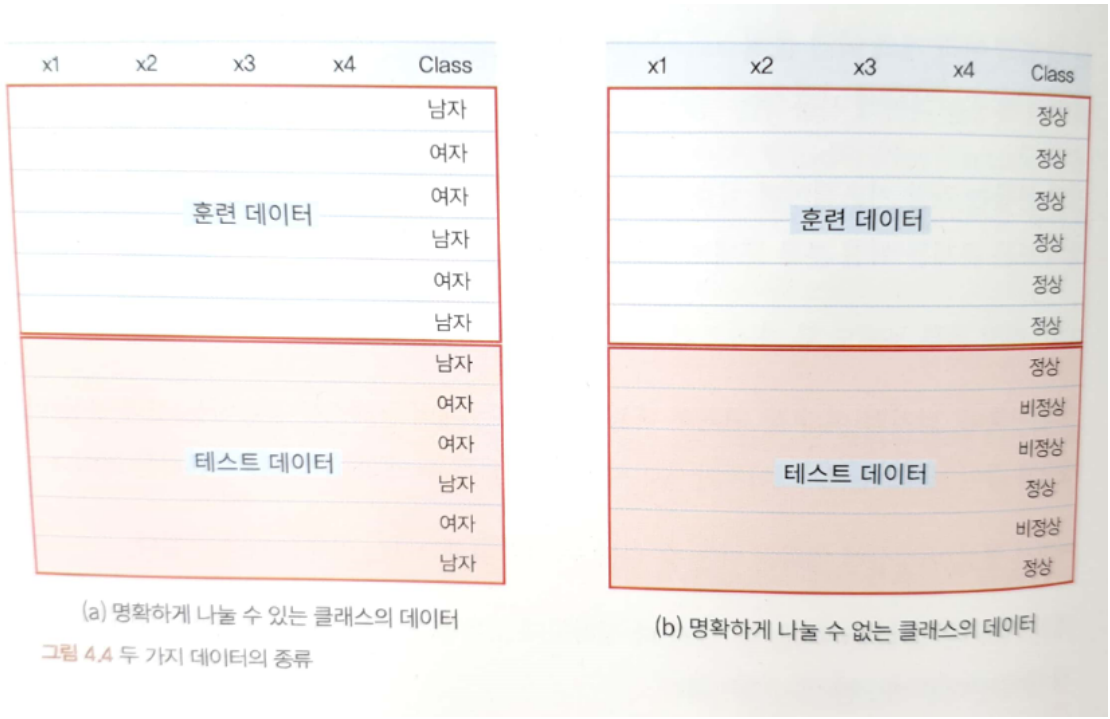


그림 4.3 정상 상태와 비정상 상태들

그림 4.3과 같이 정상 상태는 명확하게 정의가 가능하지만 비정상 상태는 너무 많은 패턴을 가지고 있어서 패턴을 정의할 수 없다. 핸드폰이 정상일 때는 모든 기능이 잘 동작하지만 비정상일 때는 어떤 부품이 어떻게 동작할지를 전부 정의할 수 없기 때문이다.

명확하게 나눌 수 있는 클래스의 데이터와 명확하게 나눌 수 없는 클래스의 데이터는 알고리즘이 사용하는 훈련 데이터도 다르다.



(a) 명확하게 나눌 수 있는 클래스의 데이터

(b) 명확하게 나눌 수 없는 클래스의 데이터

그림 4.4 두 가지 데이터의 종류

그림 4.4(a)는 명확하게 나눌 수 있는 클래스를 분류하는 알고리즘을 학습시키기 위한 훈련 데이터와 테스트 데이터이다. 훈련 데이터는 데이터를 최대한 많이 수집하는 것이 중요하며, 분류할 클래스가 남자 3명, 여자 3명과 같이 데이터를 균등하게 포함하는 것이 좋다.

결정 트리, 서포트 벡터 머신(support vector machine), 신경망 등 대부분의 분류 알고리즘이 이런 데이터를 이용해서 학습하고 테스트를 한다.

그림 4.4(b)는 명확하게 나눌 수 없는 클래스를 분류하는 알고리즘을 학습시키기 위한 훈련 데이터와 테스트 데이터이다. (a)와 달리 훈련 데이터의 클래스는 정상밖에 없는 것을 알 수 있다. 이 데이터를 이용하는 알고리즘은 정상 상태를 학습한 후에 테스트 데이터가 정상 상태와 똑같은지 아니면 다른지를 구분하는 데 특화되어 있다. 알고리즘으로는 MT법(Mahalanobis Taguchi method)이 있으며 이것은 이번 장에서 배울 마할라노비스-다구찌 시스템의 기본 알고리즘이다.

수학에서 두 점사이의 거리를 계산하는 방법은 여러가지가 있지만 대표적으로 유클리드 거리와 마할라노비스 거리가 있다.

유클리드는 중학교때 배운 것이다.

마할라노비스 거리는 데이터가 가지고 있는 확률분포, 상관관계를 고려한 거리다.

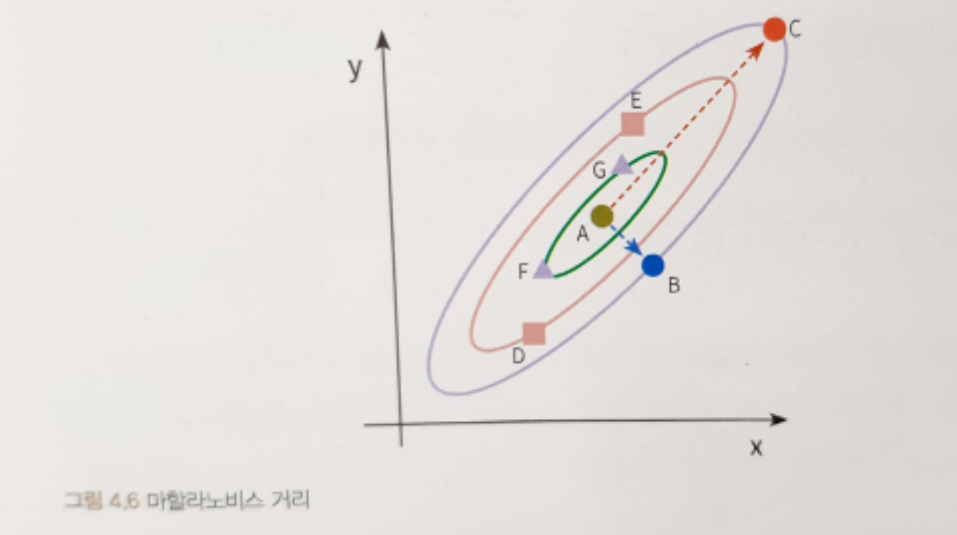


그림 4.6 마할라노비스 거리

그림 4.6에서 데이터의 중심점 A에서 점 B, 점 C와의 거리를 생각해 보자.

각 점의 좌표를 A(3,3), B(4,2), C(6,6)로 정의하면 점A와 점B, 점A와 점C의 유클리드 거리는 알아서 계산해라.

유클리드 거리로 계산하면 점C가 점 B보다 데이터의 중심 A에서 더 멀다는 것을 알 수 있다.

그러나 마할라노비스 거리는 데이터의 분포, 상관관계를 고려해 데이터 중심에서 점 A와 점 B의 거리를 계산한다.

위 그림에서 중심 A를 둘러싼 동심원 3개가 있다. 데이터 분포 관점에서 점 B와 점 C처럼 같은 색깔의 타원 위에 존재하는 점들은 데이터 중심으로부터의 같은 거리로 취급한다는 의미이다. 따라서 마할라노비스 거리 MD(A,B)로 정의하면 다음과 같다.

$$MD(A,B) = MD(A,C)$$
$$MD(A,D) = MD(A,E)$$
$$MD(A,F) = MD(A,G)$$

이 시스템을 MTS라고 하며 이는 대표적인 분류 알고리즘으로 제품의 양/불량, 시스템 등의 정상/비정상 상태를 판단하는 이상 감지 분야에서 많이 쓰이고 있을 뿐만아니라 손 글씨, 개, 고양이 등 이미지를 이용한 패턴인식 분야에서도 많이 응용되고 있다.

MTS는 다른 분류 알고리즘과 같이 모델이 필요하다. 모델이란 패턴 인식, 정상/비정상 등과 같이 시스템의 목적이 들어있는 수식이나 규칙을 의미하며, 판단하는 기준을 말한다. MTS를 이용한 이상 감지 시스템의 모델은 정상 상태의 특징을 모델이 포함하며, 모델의 출력은 정상/비정상, 양/불량과 같이 두 가지 값을 사용자에게 제시한다.

다음은 MTS를 이용한 이상감지 시스템의 개요를 보여준다.

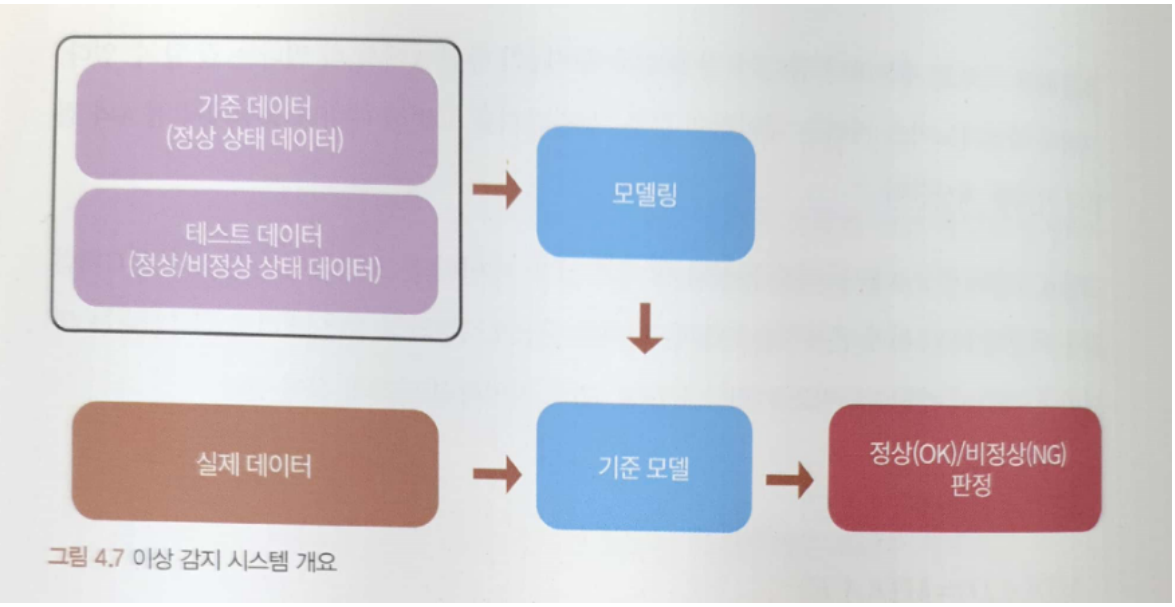


그림 4.7 이상 감지 시스템 개요

MTS는 목표 시스템의 정상 상태만을 포함하는 기준 데이터를 이용해 모델을 만들고 테스트 데이터를 이용해서 모델을 테스트 한다. 테스트 데이터는 정상 상태와 비정상 상태의 데이터가 포함되어 있으므로 모델이 정상/비정상을 잘 구분하는지를 알 수 있다. 여러 번의 검증을 통해 최적의 모델이 완성되는데 이것을 기준 모델이라고 한다. 기준 모델이 만들어지면 현장에 배치되어 실제 데이터를 가지고 정상/비정상을 판단한다.

모델링 과정에서는 시간이 걸리더라도 철저히 검증해야 한다. 그렇지 않으면 오작동하게 되어 생상라인을 멈추게 하거나, 불량을 정상으로 잘못 판단하는 경우도 발생해 큰 손실을 끼칠 수 있다.

MTS의 장점은 다른 이진 분류와 같이 정상/비정상 등 클래스를 보여주는 것 뿐만 아니라 비정상이라면 **어느정도 다른지를 수치화해 보여준다.** 이 수치화의 기준이 마할라노비스 거리가 되며, 사용자가 시스템의 고장이나 이상을 숫자로 쉽게 파악할 수 있다는 장점이 있다.

MTS의 장점은 결과의 수치화뿐만이 아니다.

다음 그림은 과거 몇 년간의 여름의 기온과 습도의 데이터 분포를 그림으로 그려낸 것이다. 기온이 높아질수록 습도는 낮아지며, 기온지 낮아지면 습도가 높아지는 반비례 관계를 볼 수 있다. 이 과거 데이터와 함께 올해 8월과 9월의 기온과 습도 데이터를 표시했다. 이상 감지 시스템은 일변량(univariate)을 이용한 이상 검출과 다변량 (multivariate)을 이용한 이상 검출로 나눌 수 있다. 일변량을 이용한 이상 검출은 한 번에 하나의 변수로 이상유무를 판단하며, 다변량을 이용한 이상 검출은 이상 유무를 두 개 이상의 변수를 동시에 사용해 판단한다. 다변량을 이용한 이상 검출 방법은 여러 가지이지만 여기서는 MTS에 대해서 설명하겠다.

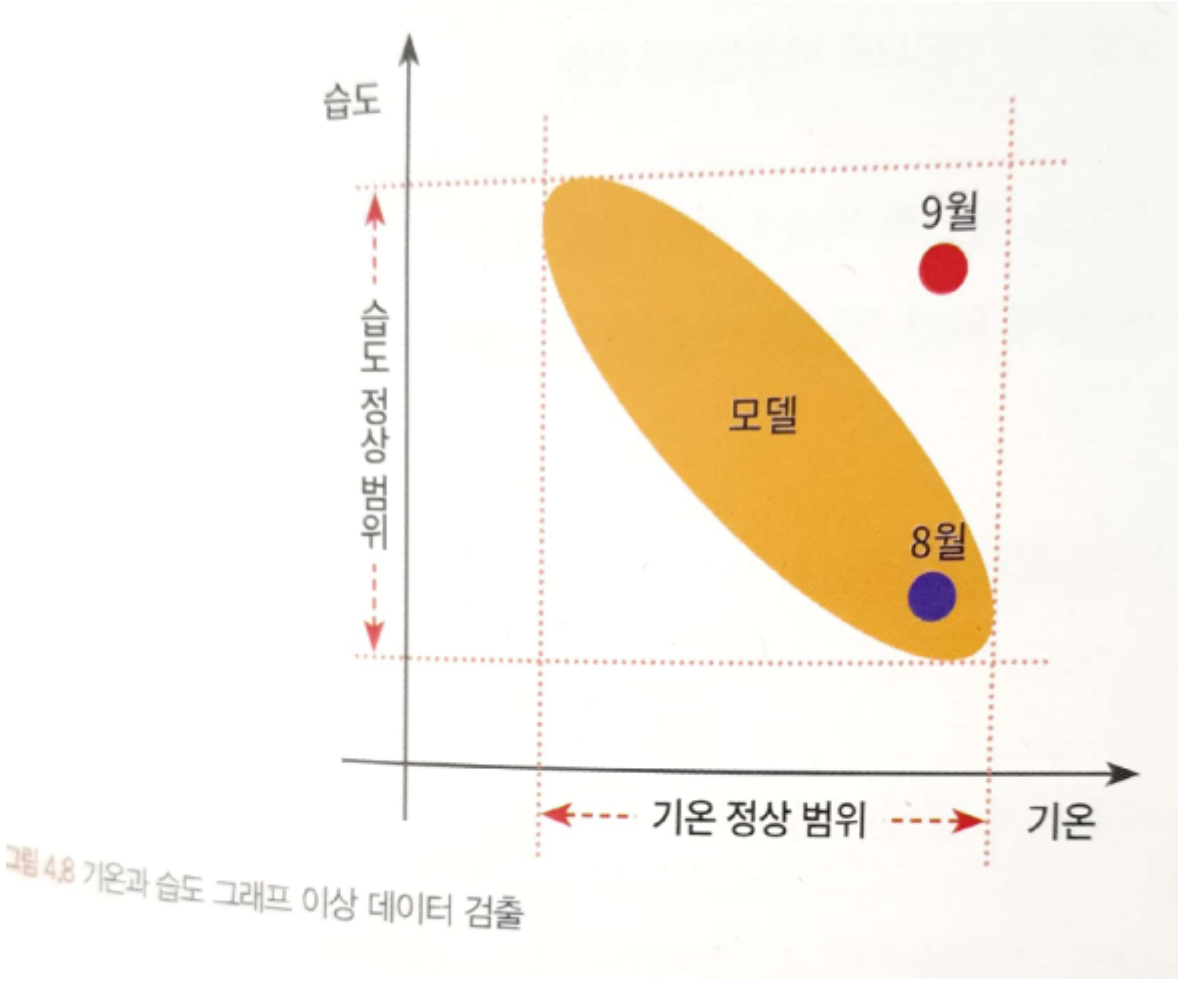


그림 4.8 기온과 습도 그래프 이상 데이터 검출

1. **일변량을 이용한 이상 검출**

일변량으로 이상을 검출할 때에는 각각의 변수에 정상 조건을 지정해서 이상을 검출하는 방식이다. 기온과 습도의 정상 범위를 각기 따로 계산해 최솟값과 최댓값을 지정해 두고 새로운 값이 정상 범위안에 있으면 정상(예년의 기온 습도 패턴과 동일한 수준으로 판정), 정상 범위 밖에 있으면 비정상(예년 기온과 습도 패턴과 다른 것으로 판정)으로 결정한다. 따라서 다음과 같이 판단할 수 있다.

- 8월 데이터 : 기온과 습도가 모두 정상 범위 안에 있으므로 **정상**
- 9월 데이터 : 기온과 습도가 모두 정상 범위 안에 있으므로 **정상**

2. MTS를 이용한 이상 검출

MTS를 이용해 이상을 검출할 때에는 데이터의 분포를 고려해 정상 조건을 지정해서 이상을 검출한다. MTS는 과거의 기온과 습도 데이터의 분포를 기준으로 위 그림과 같이 정상 범위를 모델화(오렌지색 원)한다. 그리고 새로운 값이 모델 안에 포함이 되어 있는지 여부에 의해서 정상/비정상을 판단한다.

- 8월 데이터 : 기온과 습도가 모두 정상 범위 안에 있으므로 **정상**
- 9월 데이터 : 기온과 습도가 모두 정상 범위 안에 있으므로 **비정상**

이처럼 MTS(다변량)을 이용하면 데이터의 분포를 이용하므로 좀 더 정확한 이상 판단을 할 수 있다는 장점이 있다.

초기 MTS는 역행렬을 이용해서 정상/비정상을 판단했다.

데이터 x가 2차원 (x_1, x_2)로 구성되어 있고, 각 차원의 평균이 $\bar{x} = (\bar{x}_1, \bar{x}_2,)$ 일 때, x_1, x_2의 공분산 행렬 $S_{x_1x_2}$ 은 다음과 같이 정의할 수 있다.

$$S_{x_1x_2} = \begin{bmatrix} s_{x_1x_1} & s_{x_1x_2} \\ s_{x_2x_1} & s_{x_2x_2} \end{bmatrix}$$

이 정상집단에서 어떤 점 $x = (x_{i1}, x_{i2})$ 의 마할라노비스 거리를 역행렬을 이용하면 다음과 같이 정의한다.

$$MD(x) = \sqrt{(x-\bar{x}){S_{x_1x_2}}^{-1}(x-\bar{x})`} \text{ - - - - - - - - - - - - - - - - - 식 } (4.2)$$

데이터 x가 표준화되어 있다면 공분산 행렬 $S_{x_1x_2}$는 상관행렬 $R_{x_1x_2}$ 로 바꿀 수 있다.

$$MD(z_x) = \sqrt{z_x {R_{x_1x_2}}^{-1} {z_x}`} \text{ - - - - - - - - - - - - - - - - - 식 } (4.3)$$

MTS의 마할라노비스 거리는 일반적인 마할라노비스 거리의 제곱을 사용하며, 변수의 개수 m으로 나눠준다. 이렇게 계산하면 정상 데이터로 모델링을 한 후에, 정상 데이터의 마할라노비스 거리의 평균은 1이 된다.

그리고 공분산 행렬은 앞에서 말했듯이 데이터의 단위에 따라 값이 바뀌므로 이진 분류에서 마할라노비스 거리를 사용하기에는 어려움이 있다. 그러므로 MTS에서는 상관행렬을 이용한 마할라노비스 거리를 사용한다.

따라서 MTS의 마할라노비스 거리는 다음과 같이 정의된다.

$$MD^2 = \frac{z_x {R^{-1}} {z_x}'}{m} \text{ - 식 } (4.4)$$

여기서, m은 변수의 개수이다.

예를 들어 어떤 정상그룹의 상관행렬 R과 거리를 계산하기 원하는 표준화된 데이터 x를 다음과 같이 정의할 때

$$R = \begin{bmatrix} 1 & -0.77 \\ -0.77 & 1 \end{bmatrix}, x = (0.42, 0.55)$$

계산 순서에 유의해서 마할라노비스의 거리를 구해보자.

결과는 $MD^2 = 1.03$ 이다.

이것은 정상집단의 중심에서 점 x는 1.03 만큼 떨어져 있다는 의미이다. 계산 순서는 식 4.4대로 하면 된다.

=SUMPRODUCT()는 데이터별로 전치행렬의 행렬을 곱할 때 유용하게 사용할 수 있다. 참고해라.

이렇게 역행렬을 이용한 마할라노비스 거리 MD^2 는 간단하게 계산할 수 있었다. 여기서 역행렬이 이 MTS 시스템의 모델이 된다. 모델이란 패턴 인식, 정상/비정상 등과 가이 시스템의 목적이 들어있는 수식이나 규치을 의미하며, 판단하는 기준을 말한다고 이야기 했다. 역행렬에는 정상 집단의 데이터 분포인 상관행렬 정보가 들어가 있다. 이 역행렬을 이용해서 새로운 데이터에 대한 마할라노비스 거리를 계산해서 정상/비정상의 상태를 판단하게 된다.

역행렬이 이상감지 시스템에 이용되었으나 시스템 중에는 행렬식이 계산이 안 되어 역행렬을 사용할 수 없는 케이스가 발생하기 시작했다. 그래서 역행렬을 사용하지 않아도 계산할 수 있는 MTS를 생각해야 하는 상황이 왔다.

식 (4.4)를 살펴보면 데이터의 중심이 원점인 표준화된 데이터를 역행렬로 나눠줌으로써 타원 모양의 분포가 원 모양으로 바뀌면서 중심과 점사이의 거리를 구하는 이미지 이다.

이 과정을 그른 것이 다음 그림이다.

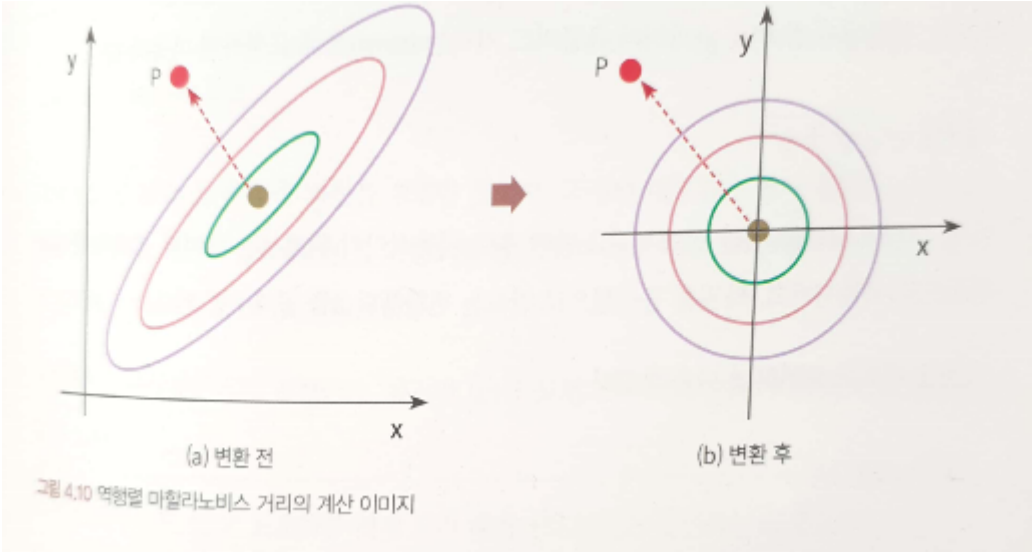


그림 4.10 역행렬 마할라노비스 거리의 계산 이미지

그림 4.10을 보자. 오른쪽 그림에서 원점에서 점 P와의 마할라노비스의 거리는 다음과 같다.

$$MD^2 = \frac{||P||^2}{2} \text{ ---------- 식 } (4.5)$$

식 4.5에서 || || 는 벡터의 크기를 나타내는 스칼라 값이며 norm(노름, 놈)이라고 한다. 놈은 일반적으로 다음과 같이 정의된다.

$$||x||_p := \sqrt{(\sum_{i=1}^{m} |x_i|p)} \geq 0 \text{ ---------- 식 } (4.6)$$

여기서, p=2가 일반적이며 다음과 같이 재정의 할 수 있다.

$$||x||_2 := \sqrt{(\sum_{i=1}^{m} |x_i|^2)} = \sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_m}^2} \text{ ---------- 식 } (4.7)$$

이제 굶 4.10 (a) 변환 전에서 (b) 변환 후로 이동시키기 위해서 데이터 분포의 중심을 원점으로 이동시키고, 타원을 동심원으로 만드는 변환행렬 A를 곱해주면 된다.

이것을 식으로 표현하면 다음과 같다.

$$P = ZA \text{ ---------- 식 } (4.8)$$

이제 식 (4.8)을 식(4.5)에 대입하면 다음과 같다.

$$MD^2 = \frac{ZAA'Z'}{m} = \frac{z_x R^{-1} z'_x}{m}$$

변환행렬 A만 구하면 역행렬을 구하지 않아도 마할라노비스 거리를 구할 수 있다.

이는 3장에서 배운 주성분분석과 동일한 과정인 것을 눈치 챌 수 있다.

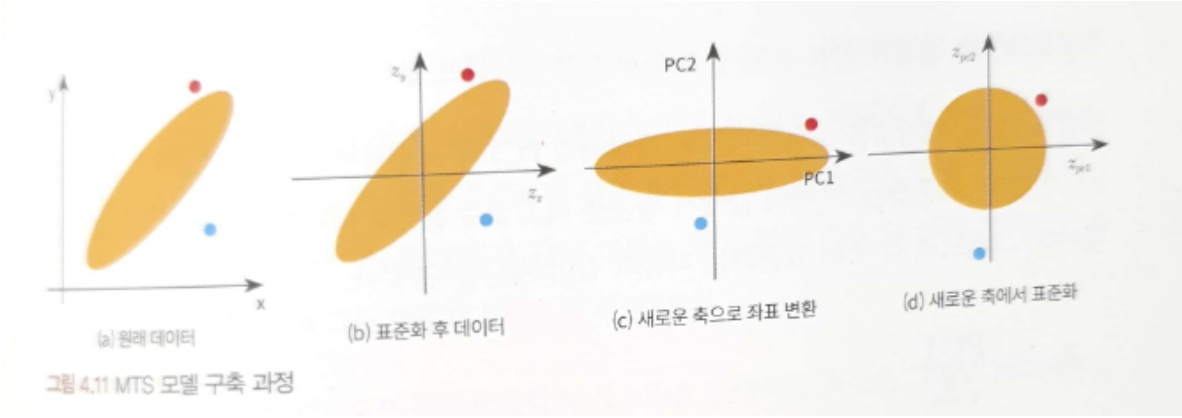


그림 4.11 MTS 모델 구축 과정

위 그림은 MTS 모델 구축 과정을 설명한 것이다. 과정의 대부분은 앞서 설명한 주성분분석과 일치한다. 데이터 분포(타원)를 기준으로 빨간 점과 파란 점은 중심에서 거의 같은 거리에 위치해 있다.

하지만 (c)를 보면 데이터의 분포가 아직 타원인 것을 알 수 있다.

(c)의 상태에서 한 번 더 표준화를 하면 데이터의 분포가 원(3차원의 경우에는 구체 모양이 된다)모양이 된다. 이 과정을 거치고 나면, 원점에서 각 점과의 유클리드 거리가 데이터의 분포를 고려한 거리 가 된다.

이 과정을 확대해 보면 다음과 같다.

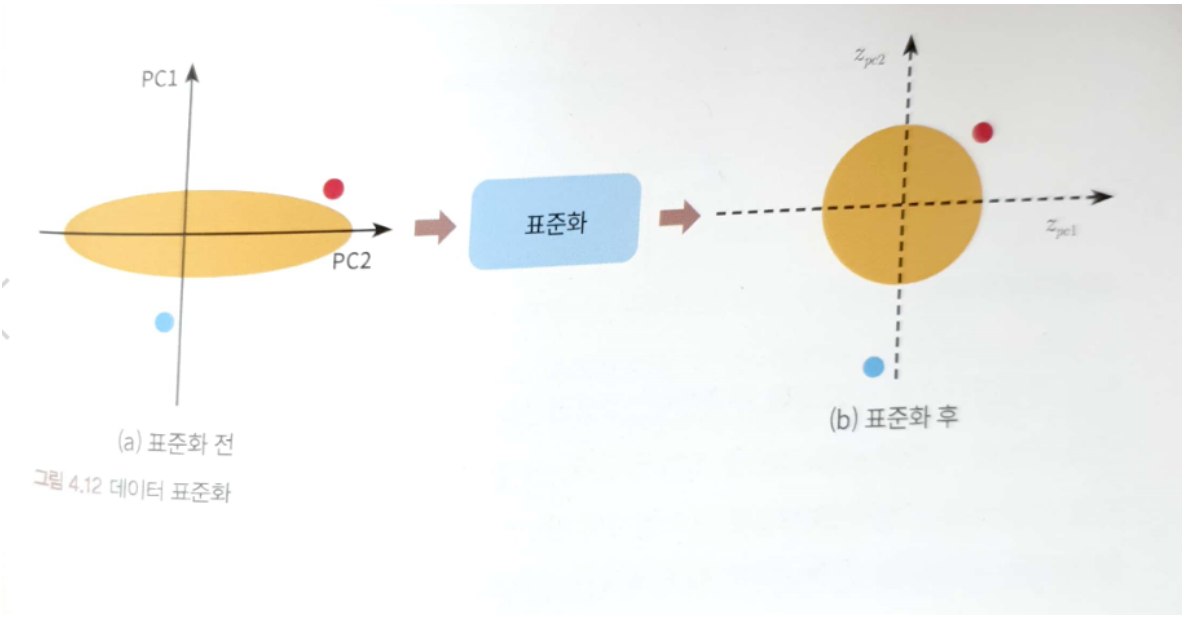


그림 4.12 데이터 표준화

주성분분석을 되돌아보자.

그림 4.12(a)에서 λ를 고윳값이라고 가정하면 PC1의 표준편차는 $\sqrt{\lambda_1}$, PC2의 표준편차는 $\sqrt{\lambda_2}$ 가 된다. 표준화는 각 축의 값에서 평균을 빼고 표준편차로 나눠주는 것을 기억하고 있을 것이다.

그러므로 각 축의 표준화는 다음과 같이 계산할 수 있다.

$$Z_{pc1} = \frac{PC1}{\sqrt{\lambda_1}}$$

$$Z_{pc2} = \frac{PC2}{\sqrt{\lambda_2}} \quad \text{식 } (4.10)$$

PC1, PC2의 평균은 0이므로 각 축을 각각의 표준편차로 나누면 PC1, PC2도 평균 0, 표준편차 1인 분포가 되므로, 그림 4.12(b) 와 같이 데이터 분포가 원형이 된다는 의미이다.

MTS모델이 구축되면 새로운 데이터를 가지고 이상 유무를 판단한다. 판단 기준은 마할라노비스 거리가 된다. 그림 4.13은 MTS를 이용한 이상 판단을 위한 구성요소와 활용 예를 보여준다.

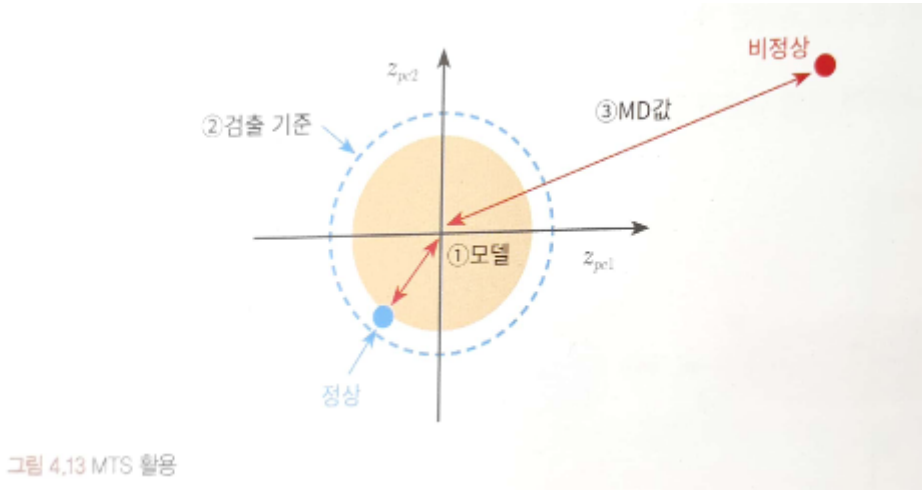


그림 4.13 MTS 활용

우선 정상집단으로 1모델을 구축한다. 정상집단을 단위공간이라고도 한다. 모델을 구축하면서 2검출(threshold)을 정한다. 이 검출 기준은 정상/비정상을 구별하는 기준으로, 간단하게는 정상집단의 분포(오렌지색 원)안쪽에 있으면 정상집단에 포함되어 있으므로 정상으로 판단하고, 정상집단 밖에 있으면 정상집단에 포함되지 않으므로, 비정상으로 판단한다.

하지만 정상집단을 포함하는 범위에서 조금 크게 검출 기준을 결정하는 것이 일반적이다. 구축된 모델을 이용해서 새로운 데이터의 3 MD값을 계산하며, 이 MD값이 검출 기준보다 작으면 정상, 검출 기준보다 크면 비정상으로 판단한다. 그림 4.13의 이미지로 이야기하면 새로운 데이터를 모델과 같이 그렸을 때 2 검출 기준 원 안에 그려지면 정상이고 검출 기준 원 밖에 그려지면 비정상으로 판단한다.

실제 데이터를 이용한 MTS를 이용한 이상감지 시스템을 구축하는 과정은 다음과 같다.

1. **모델 구축**

step_1 기준 데이터 결정

1-1. 초기 데이터로부터 이상치(outlier)를 제거한다.

step_2 기준 데이터를 이용하여 정상 모델 작성

2-1. 데이터 표준화

2-2. 상관행렬을 구한다.

2-3. 상관행렬에 대한 고윳값, 고유벡터를 구한다.

2-4. 고유벡터를 표준화한다.

2-5. 표준화된 고유벡터를 이용하여 표준화된 데이터의 좌표를 이동시킨다.

step_3 기준 데이터를 이용하여 검출 기준 결정

2. 모델 활용(이상 감지)

step_1. 구축된 모델을 이용하여 실제 데이터의 MD^2 계산

step_2. 정상/비정상 판단

먼저 모델을 구축해야 한다. 모델 구축을 크게 두 부분으로 나눌 수 있다. 첫 번째는 주성분분석을 이용해서 표준화된 주성분 공간으로 좌표를 이동시키는 것이다. 이 과정에서 모델 구축이 이뤄지며 어떤 것이 모델이 되는지는 다음 장에서 확인할 수 있다. 두번째는 구축된 모델을 이용해서 검출 기준을 결정하는 것이다.

모델 구축이 끝나면 실제 데이터를 이용해서 이상감지를 하여, 정상/비정상을 판단하면 된다.

4.2 MTS 분석 실습

2001년부터 2016년 까지 우리나라 여름의 평균기온 및 습도 데이터를 이용해 MTS를 구현해 보자.

	A	B	C
	기준 데이터		
		기온	습도
	2001년	25.00	72
	2002년	23.90	72
	2003년	23.20	74
	2004년	24.70	72
	2005년	24.40	73
	2006년	24.00	73
	2007년	24.60	70
	2008년	24.00	71
	2009년	24.20	70
	2010년	25.20	71
	2011년	24.20	73
	2012년	25.60	65
	2013년	25.90	69
	2015년	25.20	67
	2016년	26.00	67

MTS를 구현하기 위해 우선 모델을 만들어야 한다. 이 과정을 모델 구축이라고 한다.

1. **모델 구축**

   step_1 기준 데이터 결정

   1-1. 초기 데이터로부터 이상치(outlier)를 제거한다.

   <u>이상치 데이터인 2014년도를 제거 했다.</u>

   step_2 기준 데이터를 이용하여 정상 모델 작성

   2-1. 데이터 표준화

	E	F	G
	표준화 데이터		
		기온	습도
	2001년	0.42	0.55
	2002년	-1.00	0.55
	2003년	-1.90	1.35
	2004년	0.03	0.55
	2005년	-0.35	0.95
	2006년	-0.87	0.95
	2007년	-0.09	-0.24
	2008년	-0.87	0.16
	2009년	-0.61	-0.24
	2010년	0.68	0.16
	2011년	-0.61	0.95
	2012년	1.20	-2.22
	2013년	1.58	-0.63
	2015년	0.68	-1.43
	2016년	1.71	-1.43
	평균	0.00	0.00
	표준편차	1.00	1.00

2-2. 상관행렬을 구한다.

	I	J	K
		상관행렬	
		기온	습도
	기온	1	-0.76957
	습도	-0.76957	1

기온과 습도의 상관계수는 -0.7696으로 음의 상관관계이다. 빨간 원을 보면 음의 상관관계를 짐작할 수 있지만, 숫자로 정확하게 파악할 수 있다.

2-3. 상관행렬에 대한 고윳값, 고유벡터를 구한다.

2-4. 고유벡터를 표준화한다.

변환된 데이터는 아직 찌그러져 있는 것을 볼 수 있다. 유쿨리드 거리로 간단하게 계산할 수 있도록 표준화를 함으로써 데이터 공간을 동그랗게 만든다. 이 과정에서 필요한 것이 표준화이다. 그림 4.12(b)와 같이 새로운 좌표죽에서 데이터의 분포를 원으로 만드는 방법은 두 가지가 있다.

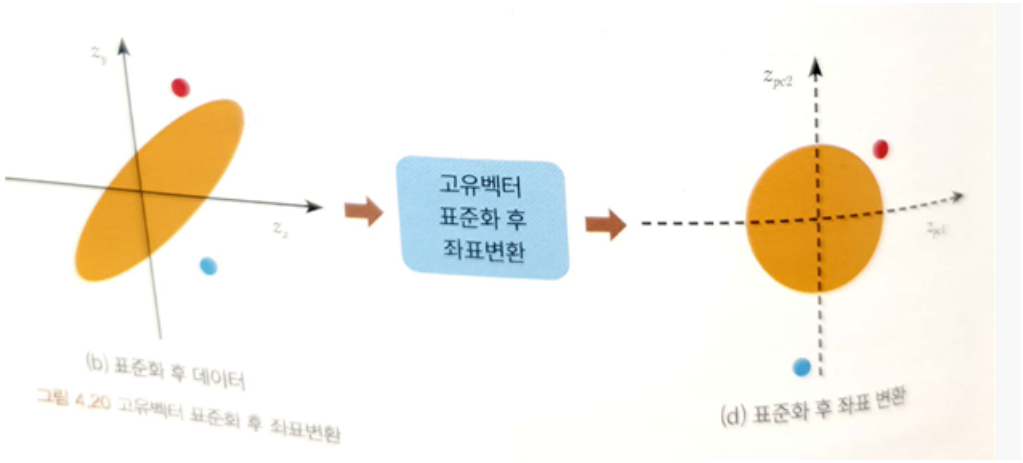
첫 번째는 좌표변환 후 표준화이다.

고유벡터를 사용해 좌표변환을 한 후에 식(4.10)과 같이 각 주성분 축에 대해 고육값으로 데이터들을 표준화하는 방법이다. 이 방법은 일반적으로 우리가 지금까지 진행해 왔던 표준화와 같은 방법이다. 이것은 그림 4.12와 같다.

두 번째는 표준화 후 좌표변환이다.

고유벡터를 표준화한 후에 좌표변환을 하는 방법이다. 표준화를 하기 위해서는 평균을 빼고 표준편차로 나눠준다. 하지만 변환된 데이터는 평균이 변함없이 0이므로, 각 데이터를 자기자신의 표준편차로 나눠주기만 하면 된다. 고윳값이 주성분의 분산과 같다는 것을 주성분분석에서 확인했다.

고윳값의 제곱근이 각 주성분 축의 표준편차와 같으므로, 고유벡터를 고윳값의 제곱근으로 나눠줌으로써 고유벡터가 표준화된다. 이것을 이미지화하면 그림 4.11이 그림 4.20과 같이 바뀐다.



첫 번째와 두 번째 방법은 순서만 다를 뿐 같은 방법이다. 이 책에서는 고유벡터를 표준화한 후에 좌표변환을 하는 두 번째 방법을 사용한다. 이렇게 표준화한 후에 좌표를 이동시키면 그림 4.20와 같이 데이터 공간이 동그랗게 변한다.

두 번째 방법인 고유벡터의 표준화를 식으로 정리해 보자.

고유벡터 V와 고윳값 λ을 다음과 같이 정의하자.

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1l} \\ v_{21} & v_{22} & \cdots & v_{2l} \\ & & \vdots & \\ v_{m1} & v_{m2} & \cdots & v_{ml} \end{bmatrix}, \qquad \lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_l\}$$

l은 주성분의 번호이며, m은 변수의 번호이다. $m = l$ 이므로 고유벡터 V는 정방행렬이다.

고유벡터의 표준화는 각 주성분의 표준편차(고윳값의 제곱근)을 나눠주면 된다.

그러므로 표준화된 고유벡터 Z_v는 다음과 같이 계산된다.

$$V = \begin{bmatrix} \frac{v_{11}}{\sqrt{\lambda_1}} & \frac{v_{12}}{\sqrt{\lambda_2}} & \cdots & \frac{v_{1l}}{\sqrt{\lambda_l}} \\ \frac{v_{21}}{\sqrt{\lambda_1}} & \frac{v_{22}}{\sqrt{\lambda_2}} & \cdots & \frac{v_{2l}}{\sqrt{\lambda_l}} \\ & & \vdots & \\ \frac{v_{m1}}{\sqrt{\lambda_1}} & \frac{v_{m2}}{\sqrt{\lambda_2}} & \cdots & \frac{v_{ml}}{\sqrt{\lambda_l}} \end{bmatrix}$$

PC1- 기온 고유벡터의 표준화를 계산해 보자.

PC1 - 기온 고유벡터(v_{11}) : -0.71

PC1의 고윳값 (λ_1) : 1.77

$$Z_{v_{11}} = \frac{-0.71}{\sqrt{1.77}} = -0.53$$

고유벡터		
	PC1	PC2
기온	-0.71	0.71
습도	0.71	0.71
고윳값		
	PC1	PC2
고유값	1.77	0.23
	1.330251	
표준화된 고유벡터		
	PC1	PC2
기온	-0.53	1.47
습도	0.53	1.47

2-5. 표준화된 고유벡터를 이용하여 표준화된 데이터의 좌표를 이동시킨다.

식 (4.11)의 표준화된 고유벡터가 기온,습도를 이용해 만든 MTS의 모델이 된다. 역행렬을 이용한 MTS에서는 상관행렬의 역행렬이 모델이라고 이야기했다. 주성분분석을 이용한 MTS에서 모델은 표준화된 고유벡터가 MTS의 모델이 된다. 표준화된 고유벡터를 가지고 MD^2 계산하고 정상/비정상을 판단한다.

이제부터는 모델의 성능 평가와 검출 기준을 결정해야 한다. 이 과정을 진행하기 위해서는 기준 데이터의 MD^2을 계산해서 MD^2의 분포를 확인하면서 결정해야 한다.

이제 우리의 모델인 표준화된 고유벡터를 이용해 좌표를 이동시키자. 새로운 축의 좌표는 다음과 같이 정의된다는 것을 주성분분석에서 확인하였다.

$$Z_{pc} = ZZ_v$$

Z :표준화된 데이터

Z_v : 표준화된 고유벡터

행렬곱을 구하는 함수 =MMULT를 이용해 새로운 축의 좌표를 구한 결과는 다음과 같다.

좌표 변환		
	PC1	PC2
2001년	0.07	1.44
2002년	0.83	-0.65
2003년	1.73	-0.82
2004년	0.28	0.87
2005년	0.69	0.88
2006년	0.97	0.12
2007년	-0.08	-0.49
2008년	0.55	-1.05
2009년	0.20	-1.25
2010년	-0.28	1.24
2011년	0.83	0.50
2012년	-1.82	-1.50
2013년	-1.18	1.40
2015년	-1.12	-1.10
2016년	-1.67	0.42

step_3 기준 데이터를 이용하여 검출 기준 결정

다시 한 번 말하지만 이 실습의 목적은 과거 기온, 습도 데이터를 이용한 이상감지 시스템 구축이다. 이상감지 시스템은 정상/비정상을 판정하기 위한 검출 기준이 있어야 한다. 그림 4.13와 같은 검출 기준을 모델 구축 과정에서 결정하여, 새로운 데이터의 값이 이 검출 기준을 만족하는지 하지 않는지를 판단하는 것이다. 이 검출 기준을 결정하기 위해서 MD^2을 구해야 한다. 주성분 공간에서의 MD^2 은 원점에서의 유클리드 거리를 계산하면 된다. 첫 번째 식(4.5)를 주성분좌표로 변형하면 다음과 같은 식이 완성된다.

$$MD^2 = \frac{||P||^2}{2} \text{ ---------- 식 } (4.5)$$

$$MD_i^2 = \sum_{j=1}^{m} {Z_{pc_{ij}}}^2 / PC\text{의 개수} \text{ ----- 식 } (4.12)$$

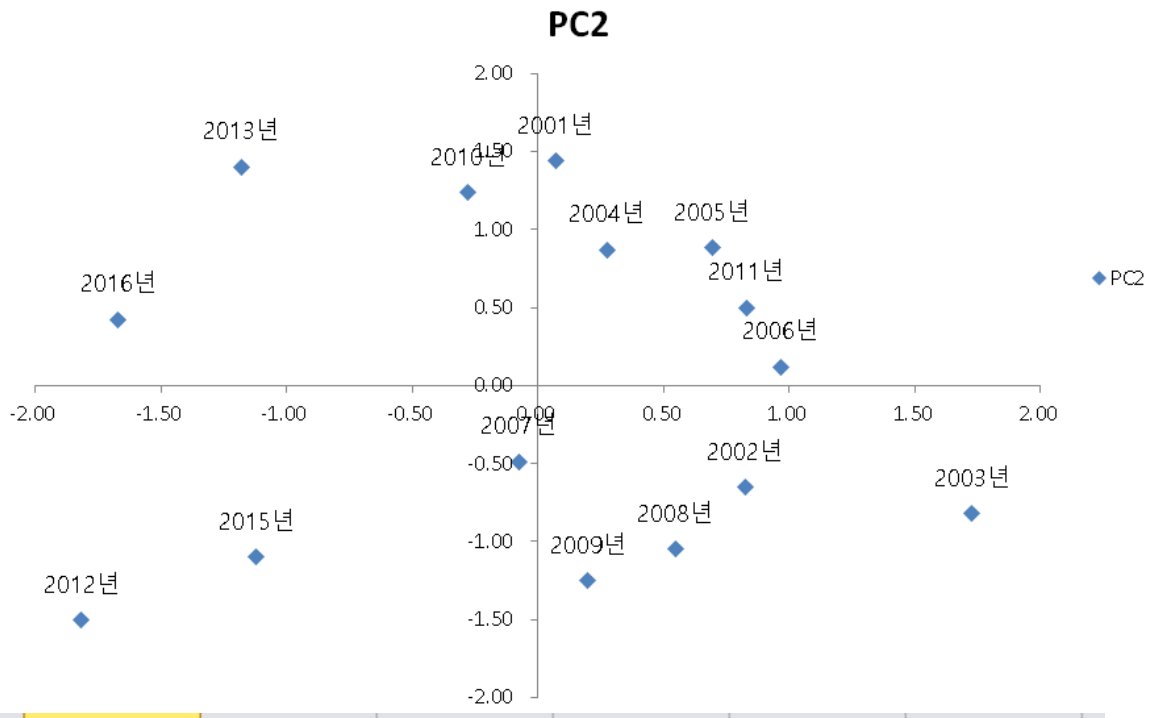
i : 데이터 번호

j: 주성분 번호

식(4.12)에서 분자는 임의의 점에서 원점까지의 육ㄹ리드 거리를 구하는 식과 같다. 이 값을 PC의 개수로 나눠주면 MD의 값이 계산된다. 2001년의 MD값을 계산해 보면 다음과 같다.

$${MD_{2001년}}^2 = \frac{0.07^2 + 1.44^2}{2} = 1.04$$

식 4.12에서 데이터의 제곱 합은 엑셀에서는 =SUMSQ()함수를 사용하면 쉽게 구할 수 있다. 말 그대로 범위 안에 있는 데이터들을 제곱해 더하는 함수이다. 한번 해보자.

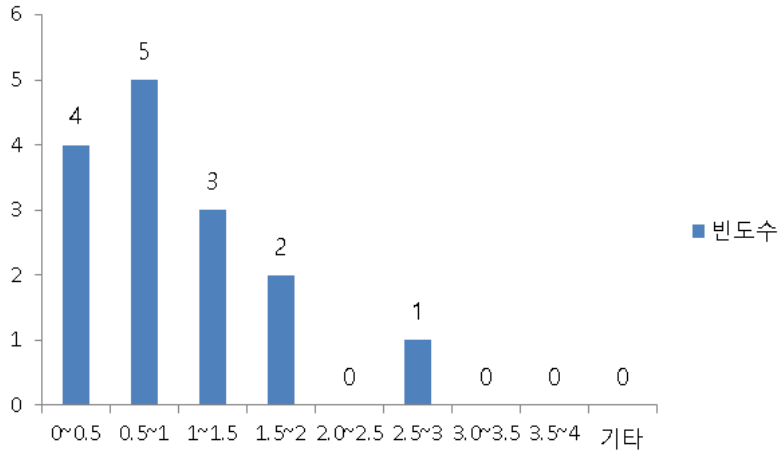


MD 계산				MD값의 통계값	
	MD			MAX	2.78
2001년	1.04	0.070474		평균	1.00
2002년	0.56			표준편차	0.671836
2003년	1.83				
2004년	0.41				
2005년	0.63				
2006년	0.48				
2007년	0.12				
2008년	0.70				
2009년	0.80				
2010년	0.80				
2011년	0.47				
2012년	2.78				
2013년	1.68				
2015년	1.23				
2016년	1.48				

MD 계산				MD값의 통계값	
	MD			MAX	2.78
2001년	1.04	0.070474		평균	1.00
2002년	0.56			표준편차	0.671836
2003년	1.83				
2004년	0.41				
2005년	0.63				
2006년	0.48				
2007년	0.12				
2008년	0.70				
2009년	0.80				
2010년	0.80				
2011년	0.47				
2012년	2.78				
2013년	1.68				
2015년	1.23				
2016년	1.48				

MD^2의 의미를 살펴보자. MD^2는 단순히 중심에서 점과의 유클리드 거리와 같다고 했다. 그림 4.23의 (a)에서 2012년 데이터가 중심에서 가장 멀리 떨어져 있고 2007년의 데이터가 가장 가깝다는 것을 알 수 있다. 아래 그림에서 MD^2를 확인해 보면 2012년의 MD^2은 2.59로 가장 크고 2007의 MD^2는 0.12로 가장작다. 이것은 2014년을 제외한 2001년 부터 2016년 우리나라 여름의 기온과 습도 데이터의 평균을 계산했을 때, 2007년이 평균에 가장 가까웠다는 의미이며, 2012년이 기준 데이터 중에서 편차가 가장 크다는 의미이다. MD^2의 통계값을 보면 앞에서 이야기 했듯이 기준 데이터의 MD^2의 평균은 1인 것을 확인 할 수 있다. 이것은 기준 데이터의 표준편차를 =STDEV.P()로 계산한 결과이며, =STDEV.S() 로 계산하면 조금 크게 나오나 깊은 의미는 없다. 이것을 기준으로 모델이 제대로 계산이 되었는지 확인할 수 있다.

기준 데이터의 MD^2의 구간별 히스토그램을 보여준다.



MD^2 히스토그램

기준 데이터의 MD^2 은 3이내에 모든 값이 들어있으며, 평균 1을 중심으로 퍼져 있는 것을 알 수 있다.

이제 MD^2의 통계값과 히스토그램을 바탕으로 검출 기준을 결정해보자.

이상 감지를 위한 검출 기준을 정하는 방법은 다음과 같이 여러 가지 방법이 있다.

첫 번째는 기준 데이터의 MD^2 값을 통해 결정하는 방법이다.

기준 데이터의 MD^2 값의 기초통계량을 보면 최댓값은 2.78이고 평균은 1, 표준편차는 0.695이다. 이 통계값을 활용해 기준을 설정하는 방법이다. 예를 들면, 기준 데이터의 모든 케이스를 포함하고 싶은 경우에는 검출 기준을 2.78로 두어 MD값이 2.78을 초과할 경우를 비정상으로 판단하도록 할 수 있다. 또 2장에서 배운 것처럼 3시그마를 기준으로 결정하고 싶을 때에는 1+3시그마 = 1+3 X 0.695 = 3.085을 검출 기준으로 설정할 수 있다. 물론 4시그마, 5시그마를 기준으로 검출 기준 삼아도 된다.

두 번째는 통계적으로 결정하는 방법이다.

MD^2는 원래의 데이터의 차원이 m이라면 MD^2은 자유도 m의 카이제곱분포를 따른다고 알려져있다. 그러나 감마분포, F분포로 가정해 자유도에 의해 기준값을 정하는 방법론이 연구되고 있다. 각 분포에 해당하는 점을 기준으로 MD^2 을 정하면 된다.

마지막으로 첫 번째와 비슷하나, 기준 데이터를 가지고 여러 가지 아이디어를 이용해서 복합적으로 판단하는 방법도 많이 쓰이며 경험적으로 3이나 4도 많이 사용된다.

여기서는 검출 기준을 3으로 정하겠다. 이렇게 되면 테스트 데이터의 MD^2 이 3이하이면 정상, 3을 초과할 경우 비정상으로 판단한다.

4.2.2. 모델 활용(이상 감지)

step_1. 구축된 모델을 이용하여 실제 데이터의 MD^2 계산

이제 모델을 이용해서 실제 데이터가 정상인지 비정상인지 결정하는 단계이다.

실제 데이터의 MD^2 는 식(4.9)를 이용하면 계산 가능하며, 변환행렬 A를 표준화된 고유벡터로 바꾸고 정리하면 다음과 같다.

$$MD^2 = \frac{ZAA'Z'}{m} = \frac{Z_{pc}Z_{pc}'}{m} \quad \text{식}(4.13)$$

$$Z_{pc} = ZZ_v \quad \text{식}(4.14)$$

Z는 표준화된 데이터, Z_v는 모델, Z_{pc}는 표준화된 주성분 공간에서의 좌표를 의미한다.

2014년을 제외한 2001년부터 2016년에 이르는 15개년 데이터를 이용해 모델을 만들고 검출 기준을 결정하였다. 기준 데이털르 이용해서 구축한 모델은 식(4.11)로 계산한 표준화된 고유벡터로, 다음 그림과 같다.

표준화된 고유벡터		
	PC1	PC2
기온	-0.53	1.47
습도	0.53	1.47

이제 실제 데이터에 대해 이상 감지가 어떻게 동작하는지 알아보자.

실제 데이터는 2017년 6월부터 9월까지의 기온, 습도 데이터이며 다음과 같다.

실제 데이터		
	기온	습도
2017년6월	23.30	62.00
2017년7월	26.90	77.00
2017년8월	25.10	71.00
2017년9월	25.60	73.00

이 데이터를 2001년부터 2016년 데이터와 비교하기 위해서는 모델을 이용해 MD^2 을 계산하고 검출기준으로 판단하면 된다. 우선 표준화를 한 후, 모델을 이용해 좌표를 이동시키고 MD^2을 구한다.

여기서 표준화할 때 주의해야 할 점이 있따. **실제 데이터를 표준화할 때에는 기준 데이터의 평균과 표준편차를 사용해서 표준화를 해야 한다는 것이다.** 그러므로 2017년 데이터는 2001~2016년 데이터의 평균과 표준편차를 가지고 표준화를 하면 된다.

2017년 6월 데이터의 기온과 습도를 표준화해보자.

기준 데이터 기온의 평균 : 24.67

기준 데이터 기온의 표준편차 : 0.77

2017년 6월 기온 : 23.3

2017년 6월 기온 표준화 = $\frac{(23.3-24.67)}{0.77}$ = -1.77

이러한 방식으로 계산한 표준화는 다음과 같다.

평균	24.67	70.6
표준편차	0.77	2.52
분산	0.5992889	6.373333

실제 데이터		
	기온	습도
2017년6월	23.30	62.00
2017년7월	26.90	77.00
2017년8월	25.10	71.00
2017년9월	25.60	73.00

평균	0.00	0.00
표준편차	1.00	1.00

실제 데이터 표준화		
	기온	습도
6월	-1.77	-3.41
7월	2.88	2.54
8월	0.55	0.16
9월	1.20	0.95
평균	0.71	0.06
표준편차	1.67	2.18

그림 4.27은 실제 데이터의 표준화를 계산한 결과이다. 다시 한 번 강조하지만 실제 데이터의 표준화는 모델을 만들었을때의 데이터(기준데이터)의 평균과 표준편차를 사용해야 한다.

그러므로 당연하지만 표준화된 실제 데이터의 평균과 표준편차도 0과 1이 아닌 것을 확인할 수 있다.

그 후 실제 데이터 표준화 4X2 행렬과 표준화된 고유벡터 2x2 행렬을 곱해서 4x2 좌표변환 행렬을 생성한다.

좌표 변환		
	PC1	PC2
6월	-0.87	-7.63
7월	-0.18	7.97
8월	-0.21	1.05
9월	-0.13	3.16

그 후, =SUMSQ 함수를 사용하여 MD계산을 한다.

step_2. 정상/비정상 판단

MD 계산		
	MD	판단
6월	29.49	비정상
7월	31.79	비정상
8월	0.57	정상
9월	5.01	비정상

그 후, 3 보다 큰 MD에 대해 비정상, 그 반대의 수치에 대해서는 정상으로 구분 짓는다.

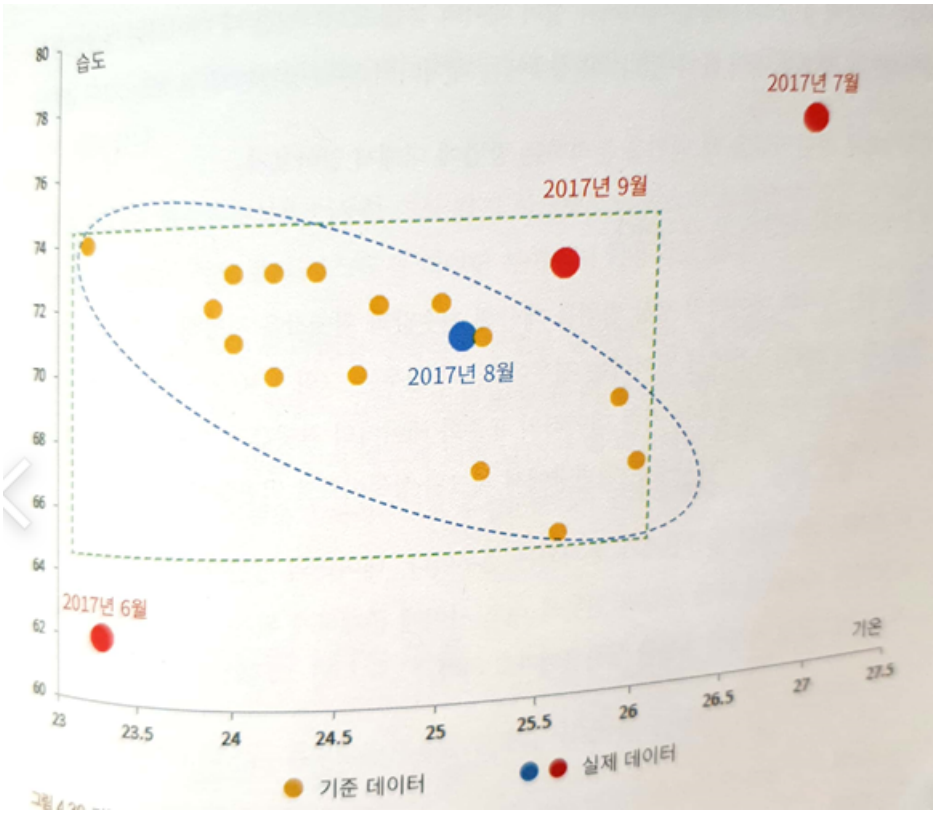
실시간 이상 감지 시스템의 경우에는 실제 데이터가 하나씩 들어오게 되므로, 6월, 7월, 8월, 9월 데이터를 따로따로 계산해서 판단해야 한다.

위 그림은 실제 데이터의 MD^2 값을 계산한 결과이다. 여기서 비정상의 의미는 2001년~2016년까지의 데이터를 기준으로 비정상이라는 의미이다.

	일변량	**MTS**
2017년 6월	비정상	비정상
2017년 7월	비정상	비정상
2017년 8월	정상	정상
2017년 9월	정상	비정상

위 표는 실제 데이터를 일변량과 MTS 판정 기준으로 판정한 결과를 비교한 표이다.

2017년 6,7,8월 데이터는 일변량과 MTS의 판정 결과가 동일하다. 하지만 9월 데이터는 일변량과 MTS의 판정 결과가 다른 것을 볼 수 있다. 왜 이런 결과가 나왔는지 알아보자.



위 그림은 기준 데이터와 실제 데이터를 그린 그래프이다. 사각형은 기온과 습도를 각 일변량으로 따로 정상 범위를 계산해 최솟값과 최댓값을 지정해 두고 이상감지를 했을 때의 정상판단조건이록 타원은 MTS로 이상감지를 했을 때의 정상판단조건이다.

실제 데이터가 정상판단조건의 내부에 위치할 때 정상으로 판정되며, 외부에 위치하면 비정상으로 판정된다.

9월 데이터를 보면 기온과 습도가 과거의 데이터 범위 (연두색 네모)에 포함되어 있어서 일변량으로는 정상으로 판정되었다. 하지만 데이터 분포(파란 타원)으로 판단했을 때에는 벗어나 있는 것을 볼 수 있으므로 MT S는 비정상으로 판정하였다. 이렇게 MTS는 데이터 분포를 이용하므로 일변량에서 보지 못하는 데이터의 분포를 기준으로 한 정상/비정상을 판단할 수 있다.

이것은 대단히 중요하다. 품질 데이터나 설비 데이터등을 모니터링할 때 데이터는 정상인데 문제가 생기는 경우를 많이 볼 수 있다.

이 경우에는 데이터의 분포(상관관계)를 보는 것이 좋다.

비정상으로 판단되었을 때 원인을 분석하는 방법에 대해서 알아보자.

비정상에 대한 원인은 두 가지 이다.

첫 번째는 각각의 데이터의 정상 범위를 계산해 최솟괎과 최댓값을 지정해 두고 이상 감지를 했을 때 정상판단조건에서 벗어나는 경우이다. 이 경우는 위 그림처럼 연두색 네모를 벗어나는 경우이다. 이 경우는 기준 데이터의 표준화 데이터의 최댓값과 최솟값을 계산해 두고 감시하면 된다. 비정상의 내용은 일변량에서 체크된 항목이므로 일변량이라고 하자.

두 번째는 데이터의 상관관계가 붕괴되는 경우이다. 데이터의 상관관계가 붕괴되면 위 그림에서 데이터가 파란 타원과 연두색 네무 사이에 존재하게 된다. 이것은 상관관계에서 체크되었으므로 비정상 내용을 상관관계라고 하자.

	기온	습도
최댓값	1.71	1.35
최솟값	-1.90	-2.22

(a) 표준화 데이터의 최대/최소값

	기온	습도	판단	비정상 내용
6월	-1.77	-3.41	비정상	일변량
7월	2.88	2.54	비정상	일변량
8월	0.55	0.16	정상	
9월	1.20	0.95	비정상	상관관계

(b) 비정상 원인 추적

그림 4.31 비정상의 원인 분석

위 그림의 (a)는 기온과 습도의 표준화 데이터의 최댓값과 최솟값이다. (b)는 이 기준을 실제 데이터에 적용해본 결과를 보여준다. 규격을 벗어난 항목에 빨간색이 칠해져 있다. 6월의 습도와 7월의 기온과 습도가 빨간색으로 되어있다. 비정상의 원인이 일변량인 것을 알 수 있다. 실제로 보면 6월의 습도는 최솟값보다 작으며, 7월의 기온과 습도는 최댓값보다 크다.

9월은 일변량으로는 감지할 수 없는 비정상을 보여준다. 이것은 상관관계가 붕괴되면서 나온 경우이기 때문이다. 예제 케이스는 변수가 두 개 여서 기온과 습도의 상관관계가 이상해지면서 나온 비정상 케이스라고 짐작할 수 있지만 변수가 많아지면 많아질수록 짐작하기가 어려워진다.

실ㅈ로, 특정 변수와 변수의 상관관계의 붕계를 명확하게 알려주는 방법은 알려진 것이 거의 없다.

이렇게 MTS는 주성분분석을 이용한 이상 감지 시스템으로 실습해 보았으묘 같은 방법으로 패턴 인식, 이상 감지 시스템 등 다양한 분야에서 활용되고 있다.

4.3 역행렬을 이용한 MTS 구축

역행렬 문제만 해결된다면, 역행렬을 이용한 방법에 기반하여 시스템을 구축하는 것도 간단하게 해결할 수 있다.

그러므로 이번에는 마지막으로 다음 데이터를 이용해서 역행렬을 이용한 사물인식 시스템을 구축해보자.

사용할 데이터는 유명한 아이리스 데이터의 클래스를 개,고양이, 토끼로 바꾸고, 변수 이름을 머리 둘레. 몸통길이, 꼬리길이로 바꿨으며 변수 한 개는 삭제한 데이터이다.

사용할 데이터는 다음과 같다.

기준 데이터 : 개			
머리둘레	몸통길이	꼬리길이	Class
5.10	3.50	1.40	dog
4.90	3.00	1.40	dog
4.70	3.20	1.30	dog
4.60	3.10	1.50	dog
5.00	3.60	1.40	dog
5.40	3.90	1.70	dog
4.60	3.40	1.40	dog
5.00	3.40	1.50	dog
4.40	2.90	1.40	dog
4.90	3.10	1.50	dog

기준 데이터 : 고양이			
머리둘레	몸통길이	꼬리길이	Class
7.00	3.20	4.70	cat
6.40	3.20	4.50	cat
6.90	3.10	4.90	cat
5.50	2.30	4.00	cat
6.50	2.80	4.60	cat
5.70	2.80	4.50	cat
6.30	3.30	4.70	cat
4.90	2.40	3.30	cat
6.60	2.90	4.60	cat
5.20	2.70	3.90	cat

앞에서 했던 것과는 다르게 기준 데이터가 두 개이다. 하나는 개를 나타내는 데이터이고, 다른 하나는 고양이를 나타내는 데이터가 10개씩 있다. 그리고 테스트 데이터를 보면 개와 고양이 그리고 토끼가 포함되어 있는 것을 확인할 수 있다.

우리가 구축할 MTS의 목적은 다음과 같다.

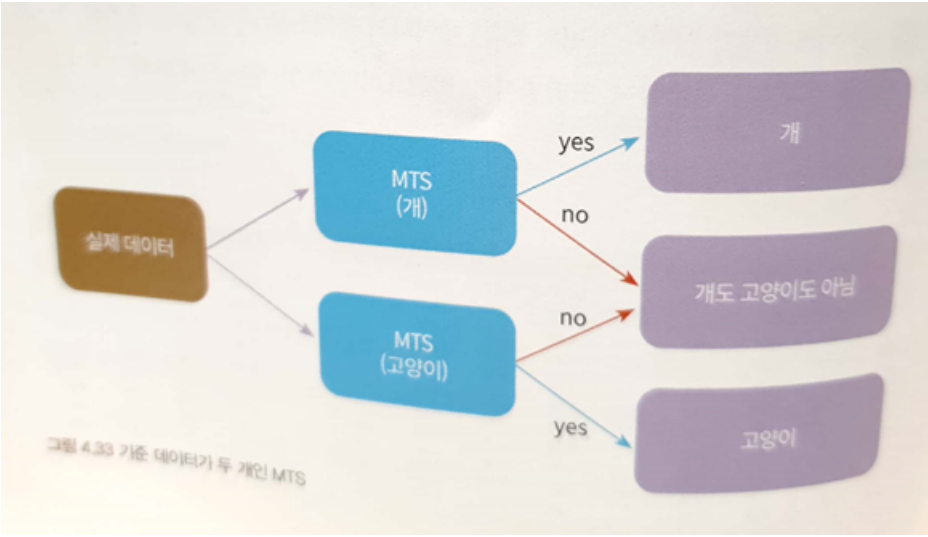
1. **개와 고양이를 구분한다.**
2. **개와 고양이가 아닌 것도 구분한다.**

목적을 잘 보면 우리는 개, 고양이, 개도 고양이도 아닌 것 세 개를 구분해야 한다.

그러나 MTS는 이진 분류여서 세 개를 구분하기는 어렵다. 불가능하다고 말하지 않고 어렵다고 말한 이유는 세 개 이상도 구현이 가능하기 때문이다. 그 중에서 단순하지만 가장 확실한 방법을 소개하려고 한다.

우선 기준 데이터가 두 개이므로 MTS가 두 개 필요하다.

두 개의 MTS는 기준 데이터 개와 고양이에 대해서 모델을 구축한다. 같은 과정을 두 번 반복한다고 생각하면 된다. 구축된 모델을 실제 데이터에 적용시킬 때, MTS(개)는 개와 개가 아님, MTS(고양이)는 고양이와 고양이가 아님을 출력한다.



위 그림이 이 모델에 대해 설명이다. MTS(개)와 MTS(고양이)가 둘 다 아니라고 했을 때에는 개도 고양이도 아닌 것이 된다.

1. 모델 구축

    step1_기준 데이터 결정

        1-1. 초기 데이터로부터 이상점(outlier)을 제거한다.

step2_역행렬을 이용하여 정상 모델 작성

2-1. 데이터를 표준화한다.

2-2. 상관행렬을 구한다.

2-3. 역행렬을 구한다.

step3_기준 데이터를 이용하여 검출 기준 결정

3-1. MD^2를 구한다.

3-2. 검출 기준을 결정한다.

2. 모듈 활용 (이상 감지)

step1_ 구축된 모델을 이용하여 실제 데이터의 MD^2을 계산한다.

step2_ 정상/비정상을 판단한다.

주성분분석을 이용한 방법과 비교해 보면, Step2 정상 모델 작성 부분만 다를 뿐 나머지는 동일하다. 개와 고양이의 기준 데이터를 다시 한 번 확인해 보자.

	기준 데이터 : 개			
	머리둘레	몸통길이	꼬리길이	Class
	5.10	3.50	1.40	dog
	4.90	3.00	1.40	dog
	4.70	3.20	1.30	dog
	4.60	3.10	1.50	dog
	5.00	3.60	1.40	dog
	5.40	3.90	1.70	dog
	4.60	3.40	1.40	dog
	5.00	3.40	1.50	dog
	4.40	2.90	1.40	dog
	4.90	3.10	1.50	dog
평균	4.86	3.31	1.45	
표준편차	0.28	0.29	0.10	
	기준 데이터 : 고양이			
	머리둘레	몸통길이	꼬리길이	Class
	7.00	3.20	4.70	cat
	6.40	3.20	4.50	cat
	6.90	3.10	4.90	cat
	5.50	2.30	4.00	cat
	6.50	2.80	4.60	cat
	5.70	2.80	4.50	cat
	6.30	3.30	4.70	cat
	4.90	2.40	3.30	cat
	6.60	2.90	4.60	cat
	5.20	2.70	3.90	cat
평균	6.10	2.87	4.37	
표준편차	0.69	0.32	0.46	

여기서 앞과 같이 모집단의 표준편차를 구하는 =STEDV.P()를 사용했다. =STEDV.S()를 사용해도 된다.

4.3.1 모델 구축

step1_기준 데이터 결정

1-1. 초기 데이터로부터 이상점(outlier)을 제거한다.

step2_역행렬을 이용하여 정상 모델 작성

2-1. 데이터를 표준화한다.

F	G	H	I
	표준화 데이터(Z)		
	머리둘레	몸통길이	꼬리길이
	0.87	0.65	-0.49
	0.14	-1.06	-0.49
	-0.58	-0.38	-1.46
	-0.94	-0.72	0.49
	0.51	1.00	-0.49
	1.95	2.02	2.44
	-0.94	0.31	-0.49
	0.51	0.31	0.49
	-1.66	-1.41	-0.49
	0.14	-0.72	0.49
평균	0.00	0.00	0.00
표준편차	1.00	1.00	1.00
	표준화 데이터(Z)		
	머리둘레	몸통길이	꼬리길이
	1.30	1.02	0.71
	0.43	1.02	0.28
	1.16	0.71	1.15
	-0.87	-1.77	-0.80
	0.58	-0.22	0.50
	-0.58	-0.22	0.28
	0.29	1.33	0.71
	-1.74	-1.46	-2.31
	0.72	0.09	0.50
	-1.30	-0.53	-1.02
평균	0.00	0.00	0.00
표준편차	1.00	1.00	1.00

표준화를 한 후에 개와 고양이의 표준화 결과가 반드시 평균 0, 표준편차 1인 것을 확인하자.

2-2. 상관행렬을 구한다.

행렬로 상관행렬을 구하는 방법은 2장에서 배웠다. 모집단의 상관행렬은 다음과 같다.

$$R = \frac{1}{n} z_x{}' z_x$$

먼저 함수 =TRANSPOSE()를 이용해서 표준화된 개와 고양이의 전치행렬 $z_x{}'$ 를 구하고, 함수 =MMULT()를 이용해서 $z_x{}' z_x$ 를 구한 후 데이터의 개수 n으로 나눠주면 된다.

2-3. 역행렬을 구한다.

이렇게 계산된 기준 데이터 개와 고양이의 역행렬이 우리가 구하려고 하는 MTS 모델이 된다.

step3_기준 데이터를 이용하여 검출 기준 결정

이제 검출 기준을 결정할 차례이다. 개와 고양이 모델인 역행렬을 이용해서 기준 데이터의 MD^2 를 계산하고, 검출 기준을 결정해 보자.

역행렬을 이용한 MD^2은 다음과 같이 계산할 수 있다.

$$MD^2 = \frac{z_x R^{-1} z_x'}{m}$$

3-1. MD^2를 구한다.

ZR-1Z'	MD2
2.35	0.78
3.96	1.32
2.35	0.78
2.63	0.88
2.37	0.79
6.74	2.25
3.78	1.26
0.34	0.11
3.23	1.08
2.25	0.75

ZR-1Z'	MD2
3.21	1.07
1.86	0.62
1.55	0.52
4.15	1.38
1.58	0.53
4.30	1.43
3.51	1.17
6.36	2.12
1.13	0.38
2.35	0.78

3-2. 검출 기준을 결정한다.

	개의 MD2	고양이의 MD2	
	0.78	1.07	
	1.32	0.62	
	0.78	0.52	
	0.88	1.38	
	0.79	0.53	
	2.25	1.43	
	1.26	1.17	
	0.11	2.12	
	1.08	0.38	
	0.75	0.78	
평균	1.00	1.00	
표준편차	0.52	0.52	
max	2.25	2.12	

이번에는 3시그마 규칙을 이용해 보자.

개와 고양이의 3시그마 규칙을 이용한 검출 기준은 다음과 같이 계산할 수 있다.

개 검출 기준 : 1 + 3 x 0.52 = 2.56

고양이 검출 기준 : 1 + 3 x 0.52 = 2.56

우연히도 개와 고양이의 검출 기준이 2.56으로 동일한 값이 나왔다. 이 예제에서 우연히 개와 고양이의 검출 기준이 같은 것일뿐, 두 개가 다른 경우도 있다.

개와 고양이 둘 다 검출 기준은 2.56으로 정하겠다. 이 검출 기준이 개의 최댓값 2.25와 고양이의 최댓값 2.12를 포함하고 있으므로 문제가 없다고 판단할 수 있다.

4.3.2 모델 활용(이상 감지)

지금까지 구축된 모델(역행렬)과 검출 기준 2.56을 이용해 실제 데이터가 어떻게 계산되고 판단되는지 알아보자.

step_1. 구축된 모델을 이용해 실제 데이터의 MD^2 계산

먼저 데이터의 표준화를 계산해보자. 앞에서도 이야기했지만, 실제 데이터를 표준화할 때에는 비교 모델의 기준 데이터의 평균과 표준편차를 이용해서 표준화를 해야 한다.

실제 데이터 표준화		
머리둘레	몸통길이	꼬리길이
1.95	1.34	0.49
-0.22	0.31	1.46
7.74	-0.38	31.72
5.57	-0.38	29.76
5.21	-0.03	44.40
3.40	-2.09	35.62
0.14	-1.06	-0.49
3.04	-1.75	25.86
5.21	-0.03	44.40

실제 데이터 표준화		
머리둘레	몸통길이	꼬리길이
-1.01	2.57	-6.20
-1.88	1.64	-5.99
1.30	1.02	0.71
0.43	1.02	0.28
0.29	1.33	3.52
-0.43	-0.53	1.58
-1.74	0.40	-6.42
-0.58	-0.22	-0.58
0.29	1.33	3.52

우리는 지금 MTS(개)와 MTS(고양이) 라는 2개 모델을 이용하고 있기 때문에 실제 데이터도 2개 모델에 각기 입력할 수 있도록 표준화해야 한다.

위 그림과 같이 개의 모델과 비교하기 위해서는 개의 기준 데이터의 평균과 표준편차를 이용해서 표준화를 해야 하며, 고양이 모델과의 비교는 고양이의 기준 데이터의 평균과 표준편차로 실제 데이터를 표준화해야 한다.

표준화가 끝났으니 $z_x R^{-1}$을 계산할 차례이다. $z_x R^{-1}$도 표준화와 마찬가지로 비교모델(역행렬)과 계산을 해야 한다. 모델을 구축할 때 계산해 두었던 역행렬을 이용해서 다음과 같이 계산을 진행해보자.

AM	AN	AO
ZR-1		
머리둘레	몸통길이	꼬리길이
2.89	-0.40	-1.04
-2.43	0.96	2.42
-1.00	-22.41	43.95
-6.02	-17.75	42.59
-19.42	-18.96	65.90
-13.86	-19.15	53.89
2.92	-3.01	-0.68
-7.88	-15.64	38.71
-19.42	-18.96	65.90
ZR-1		
머리둘레	몸통길이	꼬리길이
21.83	17.42	-40.00
16.23	15.22	-32.92
3.64	0.70	-3.17
0.37	2.14	-1.76
-16.78	-1.55	20.05
-9.98	-3.42	13.40
20.41	12.14	-34.71
-0.53	0.76	-0.71
-16.78	-1.55	20.05

이제 계속해서 MD^2 를 계산해 보면 다음과 같다. 표준화 데이터의 전치행렬이 없으므로, 앞에서 계산한 것처럼 =SUMPRODUCT()를 이용해서 데이터마다 계산한다. 마지막으로 변수의 개수 3으로 나눠주면 실제 데이터의 MD^2의 계산이 끝난다.

Q3 | =SUMPRODUCT(AI3:AK3,AM3:AO3)

AI	AJ	AK	AL	AM	AN	AO	AP	AQ
실제 데이터 표준화				ZR-1				
머리둘레	몸통길이	꼬리길이		머리둘레	몸통길이	꼬리길이		ZR-1Z'
1.95	1.34	0.49		2.89	-0.40	-1.04		4.61
-0.22	0.31	1.46		-2.43	0.96	2.42		4.37
7.74	-0.38	31.72		-1.00	-22.41	43.95		1394.62
5.57	-0.38	29.76		-6.02	-17.75	42.59		1240.94
5.21	-0.03	44.40		-19.42	-18.96	65.90		2825.77
3.40	-2.09	35.62		-13.86	-19.15	53.89		1912.36
0.14	-1.06	-0.49		2.92	-3.01	-0.68		3.96
3.04	-1.75	25.86		-7.88	-15.64	38.71		1004.61
5.21	-0.03	44.40		-19.42	-18.96	65.90		2825.77
실제 데이터 표준화				ZR-1				
머리둘레	몸통길이	꼬리길이		머리둘레	몸통길이	꼬리길이		ZR-1Z'
-1.01	2.57	-6.20		21.83	17.42	-40.00		270.74
-1.88	1.64	-5.99		16.23	15.22	-32.92		191.53
1.30	1.02	0.71		3.64	0.70	-3.17		3.21
0.43	1.02	0.28		0.37	2.14	-1.76		1.86
0.29	1.33	3.52		-16.78	-1.55	20.05		63.72
-0.43	-0.53	1.58		-9.98	-3.42	13.40		27.28
-1.74	0.40	-6.42		20.41	12.14	-34.71		192.15
-0.58	-0.22	-0.58		-0.53	0.76	-0.71		0.56
0.29	1.33	3.52		-16.78	-1.55	20.05		63.72

ZR-1Z'	MD2	판정
4.61	1.54	dog
4.37	1.46	dog
1394.62	464.87	not dog
1240.94	413.65	not dog
2825.77	941.92	not dog
1912.36	637.45	not dog
3.96	1.32	dog
1004.61	334.87	not dog
2825.77	941.92	not dog

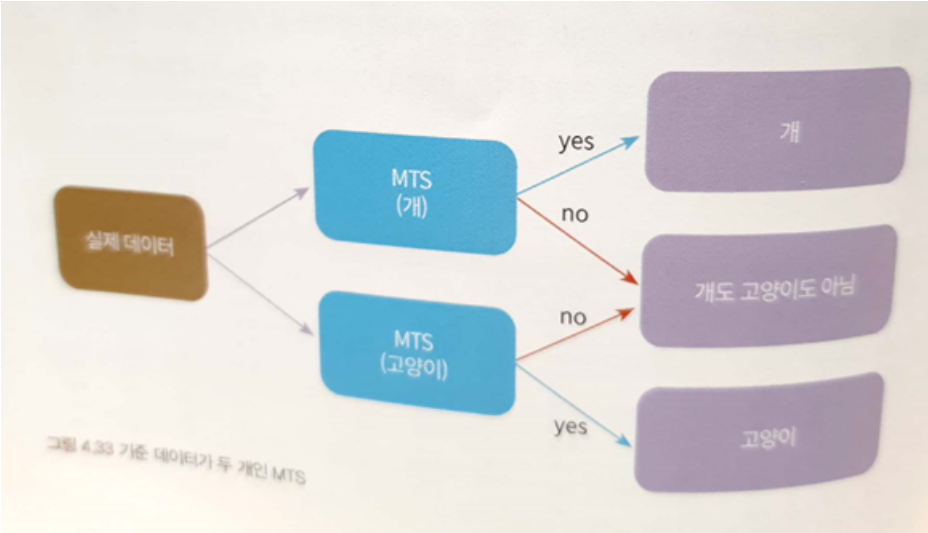
ZR-1Z'	MD2	판정
270.74	90.25	not cat
191.53	63.84	not cat
3.21	1.07	cat
1.86	0.62	cat
63.72	21.24	not cat
27.28	9.09	not cat
192.15	64.05	not cat
0.56	0.19	cat
63.72	21.24	not cat

Step1_ 정상/비정상 판단

각 모델이 계산한 MD^2을 가지고 결과를 판정해보자. 이 모델은 정상/비정상을 판정하는 주성분을 이용한 MTS와는 달리 개와 고양이인지 구별하는 시스템이다. 검출 기준 2.56을 기준을 각각 판정한 결과는 다음과 같다.

개의 MD2	판정 결과		고양이의 MD2	판정 결과
1.54	dog		90.25	not cat
1.46	dog		63.84	not cat
464.87	not dog		1.07	cat
413.65	not dog		0.62	cat
941.92	not dog		21.24	not cat
637.45	not dog		9.09	not cat
1.32	dog		64.05	not cat
334.87	not dog		0.19	cat
941.92	not dog		21.24	not cat

위 그림을 어떻게 정리하면 좋을까? 앞에서 우리가 그렸던 아래 그림을 보면 개나 고양이로 판정이 되는 경우에는 그 결과를 따르고, 개도 고냥이도 아닌 경우에는 '개도 고양이도 아님' 이라고 판정하는 것을 알 수 있다.



그러므로 이를 다시 정리하면 다음과 같다.

머리둘레	몸통길이	꼬리길이	Class	최종판정결과
5.40	3.70	1.50	dog	dog
4.80	3.40	1.60	dog	dog
7.00	3.20	4.70	cat	cat
6.40	3.20	4.50	cat	cat
6.30	3.30	6.00	rabbit	not dog, not cat
5.80	2.70	5.10	rabbit	not dog, not cat
4.90	3.00	1.40	dog	dog
5.70	2.80	4.10	cat	cat
6.30	3.30	6.00	rabbit	not dog, not cat

위 그림을 보면 개는 개로, 고양이는 고양이로 정확하게 분류한 것을 알 수 있다. 또 하나 지켜볼 것은 원래 클래스에 토끼가 있었지만, MTS는 토키에 대한 정보가 없으므로 분류 결과가 모델을 만들 때 사용했던 기준에 대해서만 판정을 하게 된다. 여기서 MTS는 적어도 개와 고양이는 아니라고 판정하는 것이다. 토끼를 판정하고 싶다면, 토끼에 대한 모델을 추가해 주면 된다.

클래스가 여러 개인 MTS를 이런 식으로 간단히 구성해 보았다.

역행렬을 이용한 MTS와 주성분분석을 이용한 MTS의 결과는 같다. 역행렬을 이용한 MTS는 간단하게 시스템을 구축할 수 있는 반면에 역행렬을 구할 수 없는 경우에는 활용할 수 없다는 단점이 있다.

하지만 주성분분석을 이용한 MTS는 역행렬을 구할 수 없는 경우에도 이상 감지 시스템 구축이 가능하며, 이 방법은 품질 공학에서 많이 쓰이는 방법으로 주성분분석과 같이 차원을 줄여서 시스템을 구축할 수도 있다.