나이브 베이즈(Naive Bayes) 분류자

특허법인 가산

특허 6팀

김윤정

목 차

KASAN
on your side

The optimal classification concept

KASAN
on your side

◆ Supervised Learning

- You know the true value, and you can provide examples of the true value.
- Classification
  - Ex1. Hit or Miss
  - Ex2. Ranking A 내지 F
  - Ex3. Type Positive or Negative
  - 이미 분류될 class를 알고 있는 상태에서, classification을 함.

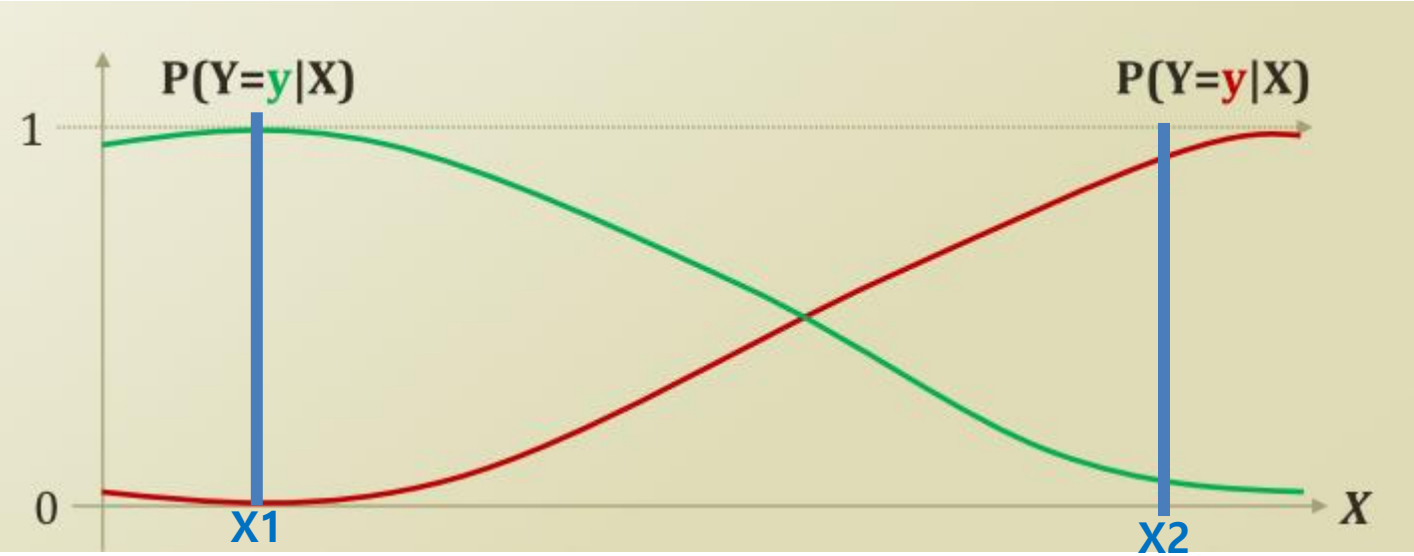
The optimal classification concept

KASAN
on your side

◆ Optimal Predictor of Bayes classifier

- Optimal predictor of Bayes classifier
  - $f^* = argmin_f P(f(X) \neq Y)$
  - Function approximation of error minimization
- Assuming only two classes of Y
  - $f^*(x) = argmax_{Y=y} P(Y = y|X = x)$

$$\sum_{y \in Y} P(Y = y|X = x) = ?$$



The optimal classification concept

KASAN
on your side

◆ Review MLE and MAP

– MLE

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{argmax}_{\boldsymbol{\theta}} \boldsymbol{P}(\boldsymbol{D}|\boldsymbol{\theta})$$

- $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
- $\hat{\theta} = \frac{a_H}{a_H + a_T}$

– MAP

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{argmax}_{\boldsymbol{\theta}} \boldsymbol{P}(\boldsymbol{\theta}|\boldsymbol{D})$$

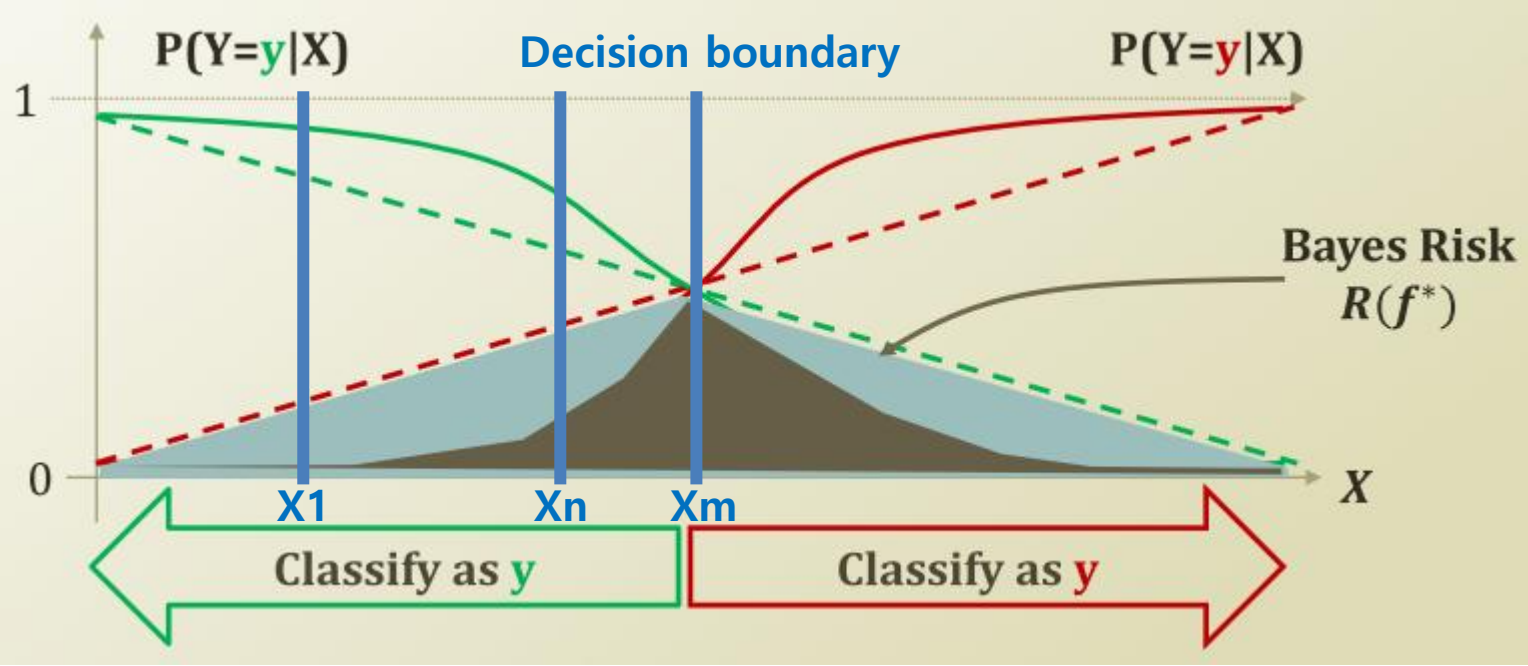
- $P(\theta|D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$
- $\hat{\theta} = \frac{a_H+\alpha-1}{a_H+\alpha+a_T+\beta-2}$

The optimal classification concept

KASAN
on your side

◆ Bayes Risk

– Optimal classification은 Bayes Risk가 최소가 되도록 한다.



The optimal classification concept

KASAN
on your side

◆ Optimal Classifier

$$f^*(x) = argmax_{Y=y} P(Y = y|X = x)$$
$$= argmax_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class Conditional Density}} \underbrace{P(Y = y)}_{\text{Class Prior}}$$

– 알아야 할 것

$$\text{Prior} = \text{Class Prior} = P(Y = y)$$
$$\text{Likelihood} = \text{Class Conditional Density} = P(X = x|Y = y)$$

– 어떻게 알 수 있는가?

- Data Set으로 부터 알 수 있음.
- 그러나, X가 여러 개의 변수를 가진다면, 변수 간의 상호작용이 이루어짐.
- 이러한 상호작용을 무시함으로써, 해결한 것이 naïve Bays classification.

The naïve Bayes classifier

KASAN
on your side

◆ Dataset

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$$

– $P(X=x|Y=y)$
$=P(x_1=\text{sunny}, x_2=\text{warm}, x_3=\text{normal}, x_4=\text{strong}, x_5=\text{warm}, x_6=\text{same}|y=\text{Yes})$
$P(Y=y)=(y=\text{Yes})$

– 얼마나 많은 파라미터가 필요한가.

$P(X=x|Y=y)$ for all x,y — $(2^d-1)k$

$P(Y=y)$ for all y — $k-1$

– d를 줄이지 않고도, 파라미터를 줄이기 위해 추가적인 가정이 필요!

The naïve Bayes classifier

KASAN
on your side

◆ Conditional Independence

- X간에 독립적임을 가정한다면?

$$P(X =< x_1, \ldots, x_i > | Y = y) \rightarrow \prod_i P(X_i = x_i | Y = y)$$

- Y가 주어졌을 때, x1과 x2가 독립적이다.

$$(\forall x_1, x_2, y) \quad P(x_1 | x_2, y) = P(x_1 | y)$$

$$P(x_1, x_2 | y) = P(x_1 | y) P(x_2 | y)$$

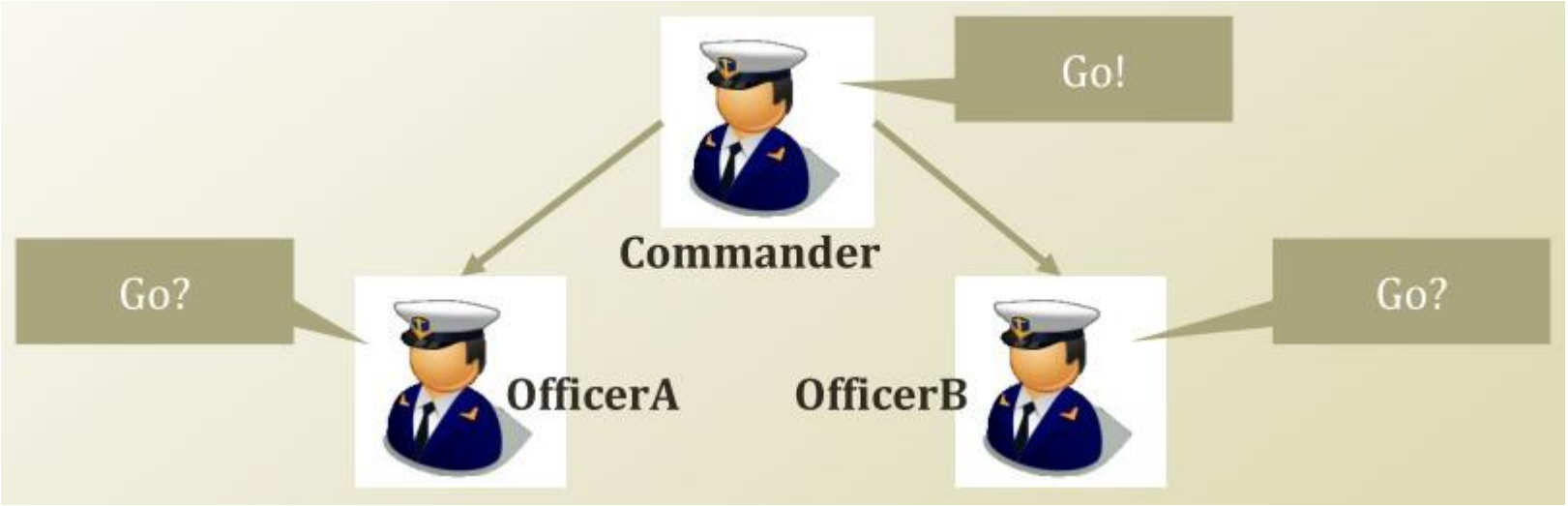
- example
  - 번개가 칠 때, 천둥이 칠 확률은 비가 오는 것과는 상관없이 같다.

$$P(\text{Thunder} | \text{Rain, Lightning}) = P(\text{Thunder} | \text{Lightening})$$

The naïve Bayes classifier

KASAN
on your side

◆ Conditional Independence



- Cf. Independence 하지 않다.

$$P(\text{OfficerA=Go}|\text{OfficerB=Go}) > P(\text{OfficerA=Go})$$

- Conditionally independent 하다.

$$P(\text{OfficerA=Go}|\text{OfficerB=Go},\text{Commander=Go}) = P(\text{OfficerA=Go}|\text{Commander=Go})$$

The naïve Bayes classifier

KASAN
on your side

◆ Dataset with Conditional Independent assumption

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

– x변수의 독립성을 가정한다면,

$$f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$$
$$\approx argmax_{Y=y} P(Y = y) \prod_{1 \le i \le d} P(X_i = x_i|Y = y)$$

– 얼마나 많은 파라미터가 필요한가.

$$\boldsymbol{P(X_i = x_i|Y = y)\ has\ (2\text{-}1)dk\ cases}$$

The naïve Bayes classifier

KASAN
on your side

◆ Naïve Bayes Classifier

– Given

- Class Prior $\mathbf{P(Y)}$
- $\boldsymbol{d}$ conditionally independent features $\boldsymbol{X}$ given the class $\boldsymbol{Y}$
- For each $\boldsymbol{X_i}$, we have the likelihood of $\boldsymbol{P(X_i|Y)}$

– Naïve Bayes Classifier Function

$$f_{NB}(x) = argmax_{Y=y} P(Y = y) \prod_{1 \le i \le d} P(X_i = x_i | Y = y)$$

- Bayes Risk 를 최소화 하는 최적의 classifier

The naïve Bayes classifier

KASAN
on your side

◆ Naïve Bayes Classifier의 문제점

- Problem1 : naïve assumption
  - 실제로 X 변수 값들은 연관성을 가진다.

- Problem2 : 부정확한 확률 계산 값Naïve Bayes Classifier Function
  - MLE 의 경우, 데이터가 충분하지 않을 경우 문제 될 수 있다.
  - 관측되지 않은 변수의 확률은 0으로 계산함.

  - MAP의 경우, stupid prior가 문제될 수 있다.
  - 그러나 prior로 MLE와 같은 문제점은 해결할 수 있다.

Apply the naïve Bayes classifier to a case study of a text mining

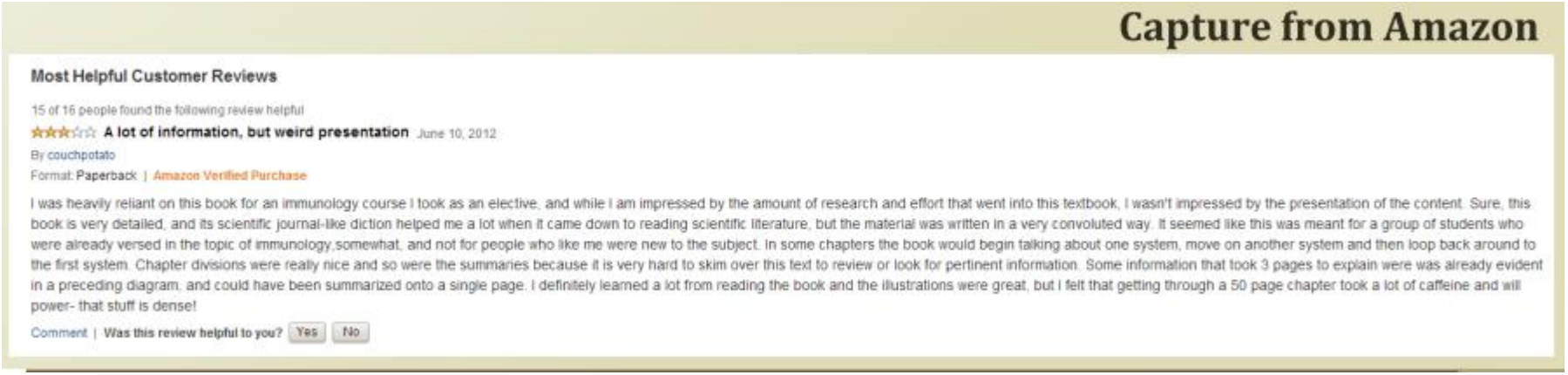
KASAN
on your side

◆ Review에 대한 Sentiment Analysis

– 리뷰가 positive 한 지, negative한지 분류하는 classifier

Capture from Amazon

Capture from Amazon



Apply the naïve Bayes classifier to a case study of a text mining

KASAN
on your side

◆ Review에 대한 Sentiment Analysis

– Hot or cool?

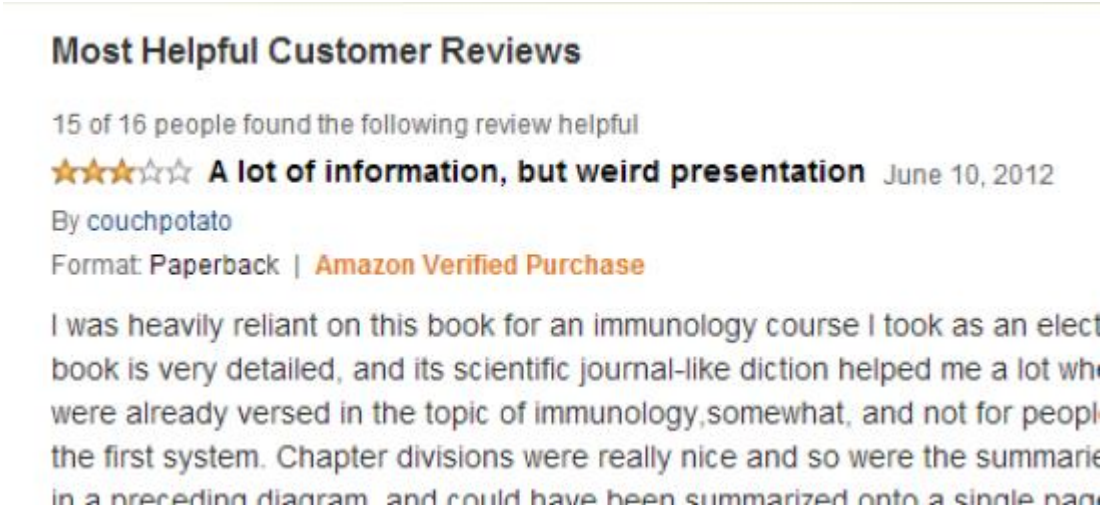
– Small or big?

Apply the naïve Bayes classifier to a case study of a text mining

KASAN
on your side

◆ Bag of words

– 리뷰 텍스트를 벡터로 변경



Most Helpful Customer Reviews

15 of 16 people found the following review helpful

★★★☆☆ **A lot of information, but weird presentation** June 10, 2012

By couchpotato

Format: Paperback | Amazon Verified Purchase

I was heavily reliant on this book for an immunology course I took as an elect
book is very detailed, and its scientific journal-like diction helped me a lot wh
were already versed in the topic of immunology,somewhat, and not for peopl
the first system. Chapter divisions were really nice and so were the summarie
in a preceding diagram, and could have been summarized onto a single pag

– 벡터 〈1,0,0,1〉
– 단어 리스트 〈I, cool, lcd, reliant〉
– 이 리뷰는 I와 reliant를 포함하고 있는 것을 알 수 있다.

Apply the naïve Bayes classifier to a case study of a text mining

KASAN
on your side

◆ Sample Dataset

- Bag of words
  - 200 documents
  - 29000 unique words
- Classes
  - Positive Sentiment
  - Negative Sentiment

- Naïve Bayes classifier에 적용

N or P

N or P가 given 인 상황에서, 개별 단어들이 등장할 확률에 대한 곱셈.

- $f_{NB}(x) = argmax_{Y=y} P(Y = y) \prod_{1 \le i \le d} P(X_i = x_i | Y = y)$
- You need to calculate…
  - $P(Y = y)$
  - $P(X_i = x_i | Y = y)$

Apply the naïve Bayes classifier to a case study of a text mining

KASAN
on your side

◆example

•File1 : bag of words

– Row: documents / col: words / value: word의 등장 여부(1or0)

	1	2	.	.	28999	29000	words
1	1	0	0	1	0	1	
2						1	
.	1	0	0	0	0	0	
.	0	0	1	0	1	1	
199	0	1	0	0	0	1	
200	0	0	0	0	1	0	
documents							

•File2 : words

1	2	.	.	28999	29000
apply	easy	.	.	simple	complicate

•File3 : document별 positive or negative 표현 (y값 given)

1	2	.	.	199	200
1	1	.	.	0	0

◆example

```
X=bagofword(sample, );
Y=sentiment(sample,:);

cntXbyY = ones(numWord,2)/1000;
for i = 1 numWord
    for j=1 N
        if X(j,i) >= 1
            cntXbyY(i,Y(j)+1)=cntXbyY(i,Y(j)+1)+1;
        end
    end
end
```

MAP 활용

```
cntY = zeros(2,1);
for j=1 N
    if Y(j) == 0
        cntY(1)=cntY(1)+1;
    else
        cntY(2)=cntY(2)+1;
    end
end

probsXbyY = zeros(numWord,2);
for i = 1 numWord
    for j=1 2
        probsXbyY(i,j) = cntXbyY(i,j) / cntY(j);
    end
end
```

MLE 활용
정규화 과정

◆example

```
probsSentiment = zeros(198,2);
for i=1:198
    for k = 1:2
        probsSentiment(i,k) = 1.
        for j=1:numWord
            if X(i,j) == 1
                probsSentiment(i,k) = probsSentiment(i,k) * probsXbyV(j,k);
            else
                probsSentiment(i,k) = probsSentiment(i,k) * (1-probsXbyV(j,k));
            end
        end
        probsSentiment(i,k) = probsSentiment(i,k) * probsV(k);
    end
end
```

Likelihood 계산

Prior 값 계산

전체 값 계산

```
logProbsSentiment = zeros(198,2);
for i=1:198
    for k = 1:2
        logProbsSentiment(i,k) = 0;
        for j=1:numWord
            if X(i,j) == 1
                logProbsSentiment(i,k) = logProbsSentiment(i,k) + log( probsXbyV(i,k) );
            else
                logProbsSentiment(i,k) = logProbsSentiment(i,k) + log( 1-probsXbyV(i,k) );
            end
        end
        logProbsSentiment(i,k) = logProbsSentiment(i,k) + log( probsV(k) );
    end
end
```

로그 함수 활용

◆example

```
estSentiment = zeros(198,1);
for i = 1:198
    if probsSentiment(i,1) > probsSentiment(i,2)
        estSentiment(i) = 0;
    else
        estSentiment(i) = 1;
    end
end
```

클래스 판별

관련 특허

KASAN
on your side

◆ 나이브 베이스 분류자와 퍼지 추론을 이용한 적조 발생 예측 방법

- 특허번호: 10-1145397
- 상태: 등록
- 특허권자: 목포대학교산학협력단

청구항 1

적조 발생시의 수온, 기온, 강수량 중 적어도 하나를 포함하는 해양환경자료를 퍼지 추론에 적합한 학습 자료로 정규화하는 전처리 단계;

상기 학습 자료를 이용하여 퍼지 추론 규칙을 생성하고, 생성된 규칙을 이용하여 입력 자료에 대해 적조 발생시의 적조생물 밀도를 예측하는 퍼지 추론 단계;

상기 학습 자료에 대한 베이스 정리를 이용하여 상기 입력 자료에 대한 분류표시의 사후 확률 값을 추정하여 적조 발생을 예측하는 나이브 베이스 분류자 단계; 및

상기 나이브 베이스 분류자 단계의 예측 결과를 기준으로 상기 퍼지 추론 단계에서 예측된 적조생물 밀도를 조정하는 후처리 단계;

를 포함하는 것을 특징으로 하는 나이브 베이스 분류자와 퍼지 추론을 이용한 적조 발생 예측 방법.

KASAN
on your side

감사합니다.

자료 관련 문의는 kyj@kspat.com 으로 주시기 바랍니다.