나이브 베이시안 학습에서 정보이론 기반의 속성값 가중치 계산방법

(An Information-theoretic Approach for Value-Based Weighting in Naive Bayesian Learning)

이 창 환 †

(Chang-Hwan Lee)

요 약 본 연구에서는 나이브 베이시안 학습의 환경에서 속성의 가중치를 계산하는 새로운 방식을 제안한다. 기존 방법들이 속성에 가중치를 부여하는 방식인데 반하여 본 연구에서는 한걸음 더 나아가 속성의 값에 가중치를 부여하는 새로운 방식을 연구하였다. 이러한 속성값의 가중치를 계산하기 위하여 Kullback-Leibler 함수를 이용하여 가중치를 계산하는 방식을 제안하였고 이러한 가중치들의 특성을 분석하였다. 제안된 알고리즘은 다수의 데이터를 이용하여 속성 가중치 방식과 비교하였고 대부분의 경우에 더 좋은 성능을 제공함을 알 수 있었다.

키워드 : 기계학습, 데이터마이닝, 나이브 베이시안

Abstract In this paper, we propose a new paradigm of weighting methods for naive Bayesian learning. We propose more fine-grained weighting methods, called value weighting method, in the context of naive Bayesian learning. While the current weighting methods assign a weight to an attribute, we assign a weight to an attribute value. We develop new methods, using Kullback-Leibler function, for both value weighting and feature weighting in the context of naive Bayesian. The performance of the proposed methods has been compared with the attribute weighting method and general naive bayesian. The proposed method shows better performance in most of the cases.

Key words : machine learning, data mining, naive Bayesian

1. 서 론

나이브 베이시안(naive Bayesian) 학습방법은 다양한 마이닝 분야에 적용되어 왔으며 알고리즘 수행의 간단함에 비하여 좋은 성능을 보여주고 있다. 하지만 나이브 베이시안의 기본적인 가정 중의 하나인 모든 속성이 같은 중요도를 가진다는 가정은 나이브 베이시안이 생성하는 이후(posterior) 확률의 정확도를 떨어지게 하는 원인으로 작용한다[1]. 예를 들어서 어떤 환자가 당뇨병이 있는지를 예측할 때, 그 환자의 혈압수치는 그 환자의 키보다 훨씬 높은 중요도를 가질 것이다. 따라서 속성별로 같은 가중치를 부여하는 것은 알고리즘을 간단하게하고 속도를 빠르게 하지만 경우에 따라서 정확도를 희생하는 경우가 있다. 따라서 나이브 베이시안에서 각 속성이 독립적이라는 가정을 변경하면 좀 더 좋은 성능을 보임을 보이는 연구도 제시되어 있다[2].

이와 같은 이유로 나이브 베이시안의 성능을 여러 방식으로 확장하는 시도들이 제안되어 왔다. 첫 번째 방식은 나이브 베이시안 학습을 속성선택(feature selection) 기능과 결합하는 방식이다[3-5]. 이 방식은 학습의 전처리 방식으로 수행되며 속성 중에서 중복되거나 학습에 의미가 없는 속성들을 제거하여서 나이브 베이시안 학습의 성능을 향상하고자하는 방식이다. 가능한 속성공간의 크기가 천문학적으로 크기 때문에 모든 속성 선택 방법의 탐색 공간의 전체를 탐색하기는 현실적으로 불가능하며 따라서 이 방법은 주어진 속성 공간에서 적절

· 본 연구는 한국연구재단(NRF)의 중견연구자 사업(과제번호: 2009-0079025)의 지원에 의하여 이루어졌음

† 종신회원 : 동국대학교 정보통신학과 교수
chlee@dgu.ac.kr

논문접수 : 2010년 5월 11일
심사완료 : 2010년 10월 4일

정보과학회논문지: 데이타베이스 제37권 제6호(2010.12)

한 속성의 집합을 찾기 위하여 주로 경험적 방식의 탐색을 이용하여 수행된다.

두 번째 방식은 나이브 베이시안의 모든 속성들에게 그 중요도에 따라 서로 다른 가중치를 부여하는 방법이다[6-8]. 속성 가중치 부여 방식은 속성선택의 방식과 연관이 있는 것으로 볼 수 있다. 속성의 가중치를 0 혹은 1 로 제한하면 속성 가중치 방식은 속성선택방식과 같아진다. 속성의 가중치를 부여함으로써 속성선택방법보다는 훨씬 유연한 방식으로 사용될 수 있으며 높은 학습 정확도를 나타낼 수 있다. 이러한 속성의 가중치 계산 방식은 주로 근접이웃(nearest neighbor) 알고리즘에서 주로 연구되어 왔다[9].

나이브 베이시안에서의 가중치 계산 방법은 근접이웃 학습 방법에 비하여 상대적으로 많은 연구가 되어 있지 않으며 최근 들어서 몇몇 방법들이 발표되고 있다. 하지만 근접 이웃알고리즘에서도 속성의 가중치부여는 알고리즘의 정확도를 향상시키는 것으로 알려져 있으며[9] 아울러 나이브 베이시안에서도 속성의 가중치 부여는 정확도를 향상시키는 것으로 알려져 있다.

본 연구에서는 나이브 베이시안 환경에서 속성의 값에 대한 가중치를 계산하는 새로운 방식을 제안한다. 기존 연구의 가중치 계산은 기존의 속성별로 가중치를 부여하는 방식인데 반하여 본 연구는 속성의 값별로 별도의 가중치를 부여하는 새로운 방식이다. 본 논문에서는 이와 같은 속성값 가중치(value weighting) 방식을 위한 새로운 가중치 부여 방법을 제안하고 이를 기존의 속성 가중치(feature weighting) 방식과 비교하여 서로의 성능을 비교하기로 한다. 나이브 베이시안에서 속성의 가중치를 부여하는 방법은 몇 가지 방법이 제안되어 있지만 속성값에 대한 가중치의 부여에 대한 방법은 아직 제안된 적이 없다. 지금까지의 방법인 속성에 대하여 가중치를 부여하는 방식을 속성기반 가중치 방식이라고 하며 본 논문의 속성 값에 대한 가중치 계산 방식은 속성값 가중치 방식이라고 이름 부르기로 한다.

2. 관련 연구

속성의 가중치 계산은 속성의 가중치로써 0과 1 사이의 실수를 부여함으로써 속성 선택의 방법보다도 좀 더 유연하며 속성선택은 속성 가중치 방법의 일부분으로 간주할 수 있다. 이러한 속성의 가중치 계산 방법은 속성에 대한 바이어스(bias)의 일종으로써 지금까지 대부분 근접이웃 알고리즘의 경우에 주로 사용되어 왔다[9]. 속성의 가중치 계산 방법은 여러 가지 방법이 제안 되어 있지만 이들은 크게 필터(filter) 방법과 래퍼(wrapper)두 가지의 종류로 구분할 수 있다. 이들 방법들은 가중치의 계산 방법과 분류학습의 상호 작용 방식에 따라서 결정된다.

첫째, 필터 방법은 속성의 가중치가 알고리즘의 수행 전에 계산되며 따라서 알고리즘의 전처리의 과정으로써 수행된다. 데이터의 특징 혹은 경험적인 측정치에 의하여 결정되는 방법으로써 대부분 정보량, 정보획득비율 등의 특정치를 사용하여 속성의 가중치를 계산한다. 두 번째의 래퍼 방식은 가설을 기준으로 가중치를 계산하는 방식인데, 먼저 각 속성에 임의의 가중치를 부여하고 이러한 가중치를 사용하여 분류학습을 진행한다. 다음으로 분류학습의 성능을 기준으로 기준의 조정한 후에 다시 분류학습을 수행한다. 이와 같은 과정을 속성 가중치의 변화가 없거나 특정한 종료 조건이 만족할 때까지 반복해서 수행한다. 즉 래퍼 방법에서 속성의 가중치는 해당 가중치를 사용하는 경우에 알고리즘의 성능이 어떤가에 좌우되며 따라서 래퍼 방법은 항상 특정 학습 알고리즘과 연관되어 결정한다.

나이브 베이시안에서의 래퍼 방법의 예는 Langley와 Sage의 SBC(selective Bayesian Classifier)를 들 수 있다[2]. 이 방법은 속성의 가중치 계산이 아닌 속성 선택에 중점을 두고 있는데, 나이브 베이시안의 정확도를 기준으로 선택된 속성집합의 성능을 평가한다. Langley와 Sage는 전체 속성에서 알맞은 속성 부분집합을 구하는 SBC라는 greedy 알고리즘을 제안하였다.

속성에 가중치를 부여하는 방법은 근접이웃 알고리즘에서는 많이 제안 되어 있으나 나이브 베이시안의 환경에서는 극히 제한적으로 연구가 진행되어 왔다. Hall[7]은 결정트리를 이용한 나이브 베이시안에서의 가중치 계산 방법을 제안하였다. 이 방법은 우선 가지치기(pruning)전의 결정트리를 생성하고 각 속성이 결정트리의 어느 레벨에서 나타나는지를 검사한다. 즉 상위 레벨에 나타나는 속성 일수록 중요도가 크다고 할 수 있으므로 해당 속성의 레벨을 기준으로 속성의 가중치를 계산한다. 결정 트리의 생성 시에 배깅(baggin) 방법을 이용하여 오차를 줄이고 결정트리를 안정화시켰으며 결정트리에 나타나지 않은 속성의 가중치는 0으로 결정하였다. 이와 같은 방법으로 기존의 나이브 베이시안에 비하여 분류의 성능이 향상됨을 보였다.

Zhang과 Sheng[8]은 나이브 베이시안에서 정보획득비율(gain ratio)을 이용한 속성 가중치 계산 방법과 아울러 다른 피드백 방법을 사용하여 속성의 가중치를 계산하였으며 성능 평가의 기준으로 AUC 기준을 적용하였다. Gartner[6]는 SVM 을 이용하여 속성의 가중치를 계산하는 방법을 제안하였다. 이 방법은 가상공간을 최적으로 이 등분할 수 있는 hyperplane을 찾아내고 이 hyperplane 을 결정하는 가중치를 나이브 베이시안에서 해당 속성의 가중치로 계산하는 방식이다. 각 가중치들

은 과적합(overfitting) 문제를 피하기 위하여 최적화되었고 해당 방법은 다른 기계학습의 알고리즘보다 좋은 성능을 보인다고 기술하고 있다.

3. 가중치 기반 나이브 베이시안 방법

분류학습에서 알고리즘은 학습 후에 새로운 데이터에 대하여 목적속성(target attribute)의 값을 부여한다. 새로운 데이터가 n개의 속성을 가지고 이들 속성의 값을 $a_1, a_2, \ldots, a_n$ 이라고 하자. 속성 C는 우리가 예측해야 하는 목적속성을 의미하고 c는 목적속성 C의 가능한 값들의 의미한다고 하면, 나이브 베이시안에서는 모든 속성들이 독립적이라는 가정에 의하여 다음의 관계가 성립한다.

$$P(a_1, a_2, \ldots, a_n|c) = \prod_{i=1}^{n} P(a_i|c) \quad (1)$$

새로운 데이터는 최대의 이후 확률값(posterior)을 가지는 속성 값을 자신의 목적속성 값으로 예측하게 된다. 다시 말해서 새로운 데이터 D의 목적속성 값은 다음의 식에 의하여 결정된다. 이 식에서 a_{ij}는 i번째 속성의 j번째 값을 의미한다.

$$V_{NB}(D) = \arg max_c P(c) \prod_{a_{ij} \in D} P(a_{ij}|c) \quad (2)$$

새로운 데이터의 목적속성 값은 최대의 사후확률 값을 가지는 목적속성의 값을 가진다.

나이브 베이시안에서 가정하는 모든 속성이 서로 독립적이라는 가정이 현실적으로 항상 만족하지는 않으므로 이러한 독립성 가정을 보완하는 연구가 진행되어 왔다. 첫 번째 방법은 속성선택(feature selection)의 방법이고 두 번째 방법은 속성의 가중치 부여(feature weighting) 방법인데, 속성선택은 속성 가중치 부여방법의 일부분으로 간주할 수 있다. 속성에 가중치를 부여한 나이브 베이시안의 방법은 아래와 같은 방법으로 목적속성의 값을 구한다.

$$V_{WNB}(D) = \arg max_c P(c) \prod_{a_{ij} \in D} P(a_{ij}|c)^{w_i} \quad (3)$$

여기서 w_i는 i-번째 속성의 가중치를 의미하며 속성의 중요도를 0과 1 사이의 숫자로 표현한다.

본 논문에서는 이와 같은 속성에 가중치를 부여하는 방법을 좀 더 세분화 하여서 속성의 값마다 가중치를 부여하는 방법을 제안한다. 그림 1은 나이브 베이시안에서 속성마다 가중치를 부여하는 것의 의미를 설명하고 있다. (그림 1의 #(Y)는 데이터의 값이 Y인 데이터의 개수를 의미한다.)

그림 1에서 목적속성의 이전(prior) 확률 분포는 P(Y)=9,000/10,000=0.9 이며 P(N)=1,000/10,000=0.1 이다. 성별 속성에서 속성의 값이 'male'인 경우에 이들의 사후

기타 속성들	성별	기타 속성들	목적속성
.....	male : #(Y)= 100 #(N)= 900 female : #(Y)= 8,900 #(N)= 100	.....	#(Y)=9,000 #(N)=1,000
.....	.....	.....	.....

그림 1 속성값 가중치의 예시

(posterior) 확률을 구하면 P(Y|male)=100/1,000=0.1 이며 P(N|male)=900/1,000=0.9 이다. 이는 목적속성의 이전 분포와 비교하면 상당한 변화를 보이고 있으며 따라서 성별의 값이 'male'인 경우는 목적속성의 분포에 상당한 영향을 미친다. 반면에 속성의 값이 'female'인 경우에 이들의 사후 확률을 구하면 P(Y|female)=8,900/9,000=0.989 이며 P(N|female)=100/9,000=0.011 이다. 이는 목적속성의 이전 분포와 비교하면 어느 정도의 변화를 보이지만 성별의 값이 'male'인 경우에 비하면 거의 미미한 변화를 보이고 있다. 따라서 성별의 값이 'female'인 경우는 목적속성의 분포에 상대적으로 미미한 영향을 미친다는 것을 알 수 있다.

즉 성별의 속성이 목적속성에 영향을 미치는 것의 대부분은 속성의 값이 'male'일 때에 발생하는 영향임을 알 수 있다. 하지만 기존의 방법들은 가중치 계산에서 속성별로 가중치를 부여하므로 이러한 현상을 표현할 수 있는 방법이 제한되고 있다. 따라서 이러한 문제점을 보완하기 위하여 본 연구는 속성의 값 별로 가중치를 계산하는 방법을 제시한다. 이와 같이 속성의 값에 대하여 가중치를 부여하게 되면 나이브 베이시안의 분류식은 다음과 같아진다.

$$V_{VWNB}(D) = \arg max_c P(c) \prod_{a_{ij} \in D} P(a_{ij}|c)^{w_{ij}} \quad (4)$$

이 식에서 w_{ij}는 i번째 속성의 j번째 값의 가중치를 의미한다.

4. 속성값 기반의 가중치 계산방법

앞에서 기술한 바와 같이 나이브 베이시안에서 속성의 가중치 계산은 몇 가지가 제안되어 있다. 하지만 지금까지 속성의 값 각각에 대하여 다른 가중치를 부여하려는 시도는 알려진 것이 전혀 없다. 본 논문은 나이브 베이시안의 가중치 계산에서 새로운 시도를 하려한다.

기존의 가중치 계산 방법들은 각 속성에 대하여 한 개의 가중치를 부여한다. 따라서 해당 속성의 속성값마다 같은 가중치를 사용하게 된다. 하지만 나이브 베이시안에서는 각 속성의 값마다 분류학습에 영향을 미치는 중요도가 다르며 따라서 이들의 가중치를 조절하여야 한다. 예를 들어서 그림 1에서 {P(Y)=0.9, P(N)=0.1, P(Y|male)=0.1, P(N|male)=0.9, P(Y|female)=0.989, P

(N|female)=0.011} 이다. 지금까지의 속성가중치 계산방법은 성별 속성에서의 'male', 'female' 속성값에 대하여 같은 가중치를 부여한다. 하지만 그림 1에서 보듯이 성별 속성의 값이 'male' 혹은 'female'인 경우에 목적속성에 미치는 영향은 많은 차이를 보인다. 속성값이 'male'의 경우에는 목적속성의 이전 값 분포와 많은 차이를 보이지만 속성 값이 'female'의 경우에는 목적속성의 이전 값 분포와 많은 차이를 보이지 않는다. 따라서 각 속성의 값마다 같은 가중치를 부여하는 기존의 방법 보다는 서로 구분된 가중치를 부여하는 것이 더욱 정확한 학습을 가능하게 할 것이다. 이와 같은 이유로 본 연구에서는 각 속성의 값에 대하여 다른 속성 가중치를 부여하는 방법을 제안한다. 속성 전체에 같은 가중치를 부여하면 이러한 속성 값에 따른 영향을 탐지할 수 없으며 따라서 전체적인 성능에 영향을 미칠 수도 있다.

앞에서 언급한바와 같이 속성의 가중치 계산 방법은 몇 가지 제안된 내용이 있지만 속성 값의 가중치 계산 방법은 지금까지 알려진 연구가 전혀 없다. 따라서 속성 값의 가중치 계산 방법은 본 연구에서 새롭게 고안하여야 한다. 또한 속성값 가중치 계산 방법과 더불어 속성 가중치의 계산 방법도 같이 개발하여 두 방법 간의 성능을 비교하려 한다.

본 논문에서는 속성 값 가중치의 계산을 위하여 정보이론을 이용한 방법을 제시한다. 정보이론의 방법은 가중치 계산 방식에 있어서 가장 많이 적용되는 방법의 하나이며 또한 가중치의 계산에 있어서도 이론적인 설명과 근거를 제공할 수 있는 방법이기 때문이다. 가중치의 계산 방식에 있어서 정보이론을 사용하는 방법이 나이브 베이시안에서 가장 우수한 속성 혹은 속성 값의 가중치를 제공할 수 있는지는 논의의 여지가 있을 수 있다. 하지만 정보이론을 이용한 속성 값의 계산은 가장 널리 사용되는 가중치 계산 방법 중의 하나이며 또한 본 논문의 주제는 속성 가중치의 방법과 속성 값 가중치의 방법을 비교하고 속성 값 가중치의 방법이 좀 더 뛰어난 성능을 보일 수 있는지를 검증하는데 주된 목적이 있다. 따라서 속성 가중치 계산 방법과 속성 값 계산 방식에 같은 방법을 적용하여 공정한 비교와 평가를 하려 한다.

본 논문에서 정보이론을 이용하여 속성의 가중치를 계산하는 주된 아이디어는 다음과 같다. 속성의 특정한 값이 관측되었을 때 그 관측치는 목적속성에 어느 정도의 정보량을 전달한다고 가정한다. 따라서 전달하는 정보의 량이 많은 속성 혹은 속성의 값은 많은 중요도를 가진다고 생각할 수 있다. 따라서 특정한 속성 값이 관측될 때 이들이 전달하는 정보의 량을 측정하는 것이 중요한 요소가 된다.

본 논문에서는 이러한 정보의 양을 목적속성의 이전 확률 분포와 이후 확률 분포의 차이로 정의한다. 즉, 속성 값이 관측되기 전의 목적속성의 확률 분포에 대하여 속성 값이 관측된 후의 목적속성 값이 어떻게 변하는지를 관측하고 이러한 확률분포의 변화 정도를 정보의 량으로 정의한다. 그러면 어떠한 방식으로 두 확률분포의 차이를 정의할 것인가? 이러한 목적으로 가장 널리 사용되는 방법은 정보획득(information gain) 공식이다. 정보획득 함수는 아래와 같이 정의되며 이는 C4.5[10] 등에서 분기노드(branching node)의 결정에 사용되며 가장 널리 사용되는 엔트로피 함수의 하나이다. 목적속성을 C라 하고 그 값을 c라 하면 특정한 속성 A의 속성값 a가 가지는 엔트로피 함수는 다음과 같이 정의된다.

$$H(C)-H(C|A)= \sum_a P(a)\sum_c P(c|a)\log P(c|a)-\sum_c P(c)\log P(c) \quad (5)$$

이와 유사한 정보함수를 사용하는 시스템으로서는 법칙생성 시스템인 CN2[11]를 들 수 있는데 CN2는 법칙의 생성을 위하여 $H(C|A=a)$의 함수를 사용한다. 하지만 이는 목적속성의 이후 확률만을 고려하는 함수이기 때문에 이전 확률자체가 이미 한쪽으로 치우친(skewed) 확률분포인 경우는 의미있는 법칙을 생성하지 못하게 된다.

$IG(C|a)$를 $A=a$라는 관측치가 목적속성에 제공하는 정보의 량이라고 가정하자. 본 논문에서 정의되는 정보획득의 함수와 C4.5 등에서 정의되는 정보획득의 함수는 차이가 있다. C4.5 에서는 속성 전체에 대한 정보획득의 양을 계산하므로 각 속성값의 정보량에 대한 평균값을 구한 것이 C4.5 에서의 정보획득양이 된다. 반면에 본 논문에서는 속성 전체가 아닌 속성 값에 대한 정보획득량을 구하므로 정보획득 함수의 정의가 다소 차이를 보인다. 본 논문의 $IG(C|a)$ 는 다음과 같이 정의된다.

$$IG(C|a)=\sum_c P(c|a)\log P(c|a)-\sum_c P(c)\log P(c) \quad (6)$$

하지만 식 (6)은 목적속성의 확률분포의 차이를 측정한다는 관점에서 한 가지 문제점을 포함하고 있다. 예를 들어서 n 개의 값을 가진 속성에서 특정한 값 하나가 거의 1 에 가까운 확률을 가지고 나머지는 동일한 값을 가진 아주 작은 값이라고 가정하자. 이 경우에 이후 확률에서 다른 속성 값이 1에 가까운 값을 가지고 나머지 확률은 동일하다면 이 경우 $IG(C|a)$ 수식은 이 차이를 판단하지 못하고 0 의 값을 리턴하게 된다. 예를 들어서 앞서 그림 1에서 $IG(C|male)$ 의 값을 계산하는 경우에

$$[P(y|male)\log P(y|male)+P(n|male)\log P(n|male)] -[P(y)\log P(y)+P(n)\log P(n)]$$
$$=[0.1*\log(0.1)+0.9\log(0.9)]$$

$$-[0.9\log(0.9)+0.1*\log(0.1)]=0$$

이와 같이 $IG(C|male)$ 은 male 이라는 관측이 목적속성 C에 많은 영향을 주어도 그 값이 0 이 됨을 알 수 있다.

따라서 본 논문에서는 이러한 문제를 해결하기 위하여 다른 정보함수를 이용한다. 본 논문에서는 Kullback-Leibler[12] 함수를 이용하여 목적속성의 이전 확률과 이후 확률의 차이를 측정하고자 한다. 속성 i 의 j 번째 속성값 a_{ij}가 목적속성 C에 전달하는 Kullback-Leibler 측정치의 함수를 $KL(C|a_{ij})$라고 하면 $KL(C|a_{ij})$는 다음과 같이 정의된다.

$$KL(C|a_{ij})=\sum_{c}P(c|a_{ij})\log\frac{P(c|a_{ij})}{P(c)} \tag{7}$$

$KL(C|a_{ij})$ 함수는 이벤트 c와 a 사이의 정보량을 목적속성 C의 관점에서 평균을 구한 것이다. 이 함수는 0 혹은 양수의 성질을 가지고 있으며 반대로 $IG(C|a)$는 음수와 양수의 값을 모두 가질 수 있다. 이 함수의 값이 0 이 되는 경우는 두 가지 확률분포가 동일한 경우에만 발생한다. 이러한 이유로 $KL(C|a_{ij})$ 함수는 두 가지 확률 분포의 차이를 측정하는 데에 적절한 함수로 판단된다. 따라서 특정한 속성값 a_{ij}에 대하여 그 속성 값의 가중치는 다음과 같이 정의 된다.

$$w_{ij}=\frac{1}{Z}\sum_{c}P(c|a_{ij})\log\frac{P(c|a_{ij})}{P(c)} \tag{8}$$

여기서 Z는 정규화 상수로서 다음의 값을 가진다.

$$Z=\frac{1}{|a_{ij}|}\sum_{\forall a_{ij}}KL(C|a_{ij}) \tag{9}$$

본 논문에서 정규화 상수 Z는 다음의 조건을 기준으로 상수가 설정되었다.

$$\sum_{\forall a_{ij}}w_{ij}=|a_{ij}| \tag{10}$$

이와 같은 w_{ij}의 가중치는 다음과 같은 값의 특징을 가짐을 알 수 있다.

아래의 정리들은 본 연구에서 제시하는 속성값 가중치 방법이 어떠한 경우에 기존의 속성 가중치 방법과 동일한 의미를 가지는지를 설명하고 있다. 즉, 아래 정리에서 설명하는 구체적인 조건들을 만족하는 경우는 속성값 가중치 방법은 속성 가중치 방법과 동일한 모델이 되므로 학습의 능력에 있어서 차이가 없어진다.

정리 1. w_{ij}와 w_{ik}가 각각 속성값 a_{ij}와 a_{ik}의 가중치라고 하면, 다음의 조건이 만족할 때 w_{ij}와 w_{ik}는 동일한 값을 가진다.

(1) $IG(C|a_{ij})=IG(C|a_{ik})$

(2) $E|_{P(c|a_{ij})}[\log(C)]=E|_{P(c|a_{ik})}[\log(C)]$

이때 $E|_{P(c|a_{ij})}[\log(C)]$은 a_{ij}의 이후 확률을 기준으로 구한 $\log(C)$의 평균을 의미한다.

증명. $IG(C|a_{ij})=IG(C|a_{ik})$에 의하여 다음 식이 만족한다.

$$\sum_{c}P(c|a_{ij})\log P(c|a_{ij})=\sum_{c}P(c|a_{ik})\log P(c|a_{ik}) \tag{11}$$

$E|_{P(c|a_{ij})}[\log(C)]=E|_{P(c|a_{ik})}[\log(C)]$ 에 의하여

$$\sum_{c}P(c|a_{ij})\log P(c)=\sum_{c}P(c|a_{ik})\log P(c). \qquad \square$$

정리 2. 위의 정리 1의 특별한 경우로서 모든 c 값에 대하여 $P(c|a_{ij})=P(c|a_{ik})$가 성립하면 $w_{ij}=w_{ik}$가 성립한다. 이는 정리 1의 일부분에 속하는 경우이며 따라서 증명을 생략한다.

5. 속성 기반의 가중치 계산방법

앞 절에서는 정보함수를 이용한 속성 값의 가중치 계산 방법을 새롭게 제시하였다. 본 논문의 주제가 속성에 가중치를 주는 방법과 속성의 값마다 다른 가중치를 주는 방법의 차이를 분석하는 것이 주된 목적이므로 본 절에서는 앞 절의 속성 값 가중치 계산에서 사용한 정보함수와 동일한 방법을 이용하여 속성의 가중치를 계산하는 방법을 제시하고자 한다. 이러한 방법을 사용함으로써 속성 기반의 가중치 계산 방법과 속성값 기반의 가중치 방법을 더욱 정확하게 비교할 수 있을 것이다.

본 논문에서 속성의 가중치 계산 방법은 앞의 속성 값 가중치를 계산할 때 사용한 $KL(C|a_{ij})$ 함수를 사용하며 이들 간의 속성 값에 대한 가중 평균값으로 정의한다. 속성의 가중치에 대한 정의는 아래와 같다.

$$w_{i-avg}=\sum_{j|i}\frac{\#(a_{ij})}{N}KL(C|a_{ij}) \tag{12}$$

이 정의의 문제점은 속성 값의 수가 많을수록 이 식의 값이 비례해서 증가한다는 점이다. 따라서 이 문제를 해결하기 위하여 가중치를 스플릿 정보(split information)[10]의 수식으로 나누어서 이러한 효과를 상쇄시킨다. 따라서 속성 가중치의 최종적인 정의는 다음과 같다.

$$w_i=\frac{\sum_{j|i}\frac{\#(a_{ij})}{N}KL(C|a_{ij})}{split} \tag{13}$$

여기서 $split$ 은 다음과 같이 정의된다.

$$split=-\sum_{j|i}\frac{\#(a_{ij})}{N}\log\frac{\#(a_{ij})}{N}=\sum_{j|i}\frac{\#(a_{ij})}{N}\log\frac{N}{\#(a_{ij})} \tag{14}$$

아래의 정리 3은 속성 값의 가중치와 속성 가중치의 관계를 보여준다.

정리 3. 어떤 속성 i의 모든 속성값 j에 대하여 $\#(a_{ij})$ 값이 동일하면 $w_i=\frac{1}{Z}\sum_{j|i}w_{ij}$의 관계가 성립한다.

이때 $Z_i = \frac{Z}{|a_i||\log|a_i|}$ 를 의미한다.

증명. 다음의 관계가 성립한다.

$$w_i = \frac{\sum_{j|i} \frac{\#(a_{ij})}{N} KL(C|a_{ij})}{\sum_{j|i} \frac{\#(a_{ij})}{N} \log \frac{N}{\#(a_{ij})}} = \frac{\sum_{j|i} \frac{1}{|a_i|} Zw_{ij}}{\sum_{j|i} \frac{1}{|a_i|} |\log|a_i|} = \frac{Z\sum_{j|i} w_{ij}}{\sum_{j|i} |a_i||\log|a_i|}$$

□

6. 실험 결과

본 연구에서 제안한 속성값 가중치 기반의 나이브 베이시안의 성능을 속성 가중치의 나이브 베이시안의 성능과 비교하기 위하여 다수의 데이터를 이용하여 실험을 진행하였다. 본 연구에서는 UCI machine learning [13]에서 6개의 데이터를 사용하여 실험을 진행하였다. 그림 2는 본 실험에 사용된 데이터의 특징을 설명하고 있다. 속성 혹은 속성 값의 가중치를 계산할 때에는 라플라스 스무딩(Laplace smoothing) 의 방식으로 확률을 계산하였다. 이는 각 가중치에서 사용하는 확률을 계산할 때 분모가 0이 되는 경우를 방지하기 위하여 사용하며 본 논문에서 사용하는 라플라스 스무딩 방법은 다음과 같다.

데이터	속성수	목적속성값의 수	데이터의 개수
Breast	9	10	683
Car	6	2	446
Dermatology	33	6	366
Hepatitis	3	2	320
Lung Cancer	56	2	32
Vote	16	2	232

그림 2 실험 데이터의 특징

$$P(d|a_{ij}) = \frac{\#(a_{ij} \wedge c) + 1}{\#(a_{ij}) + L}, \quad P(c) = \frac{\#(c) + 1}{N + L} \qquad (15)$$

여기서 L 은 목적속성의 값의 수를 의미하며 N 은 전체 데이터의 수를 의미한다.

본 연구에서는 속성의 값에 대하여 가중치를 계산하므로 속성의 값이 연속값인 경우는 이를 구간으로 구분하는 이산화과정이 필요하다. 본 연구에서는 이를 위하여 연속형 값들은 엔트로피 방법[14]을 이용하여 이산화(discretize)하였다. 각 데이터에 대하여 학습을 위하여 임의로 선택된 70%를 학습 데이터로 사용하고 나머지 30%를 테스트 데이터로 사용하였다. 이와 같은 과정을 5회 반복하여 평균의 값을 데이터에 대한 알고리즘의 정확도로 계산하였다.

본 연구 알고리즘들의 성능을 평가하기 위하여 각 데이터에 대하여 기본 나이브 베이시안 방법(NB), 제5장의 속성 가중치 방법(WNB), 제4장의 속성값기반 가중치 방법(VWNB)을 독립적으로 수행하여서 성능을 비교하였다. 실험은 데이터의 샘플링을 바꾸면서 5회 반복 수행하였고 알고리즘의 정확도들에 대한 평균과 범위를 그림 3에서 요약하고 있다.

데이터	NB	WNB	VWNB
Breast	81.9 ± 1.48	81.4 ± 1.69	81.6 ± 1.52
Car	91.4 ± 0.48	91.0 ± 0.52	93.2 ± 0.50
Dermatology	99.1 ± 0.05	98.8 ± 0.24	98.1 ± 0.36
Hepatitis	63.7 ± 0.77	62.3 ± 0.84	62.6 ± 0.91
Lung Cancer	71.9 ± 8.41	76.1 ± 4.27	78.3 ± 7.35
Vote	87.6 ± 0.84	87.7 ± 0.46	88.2 ± 0.62

그림 3 알고리즘들의 정확도

그림 3에서 보는바와 같이 속성 값을 기반으로 하는 나이브 베이시안 방법은 전체 6개의 데이터 중에서 5개의 데이터에 대하여 속성 기반의 나이브 베이시안 방법보다 좋은 성능을 보이고 있다. 나머지 한 개의 데이터에 대해서도 거의 같은 수준의 성능을 보여주고 있다. 따라서 속성 값을 기반으로 하는 가중치의 계산 방식은 속성을 기반으로 가중치를 계산하는 방식에 비하여 대체로 높은 성능을 제공함을 알 수 있다.

흥미로운 사항은 어떠한 가중치 방식도 사용하지 않는 일반적인 나이브 베이시안의 방식이 본 연구의 방식과 비교하여 3개의 데이터에 대하여 높은 성능을 보이고 있다. 또한 일반적인 나이브 베이시안의 방식은 속성 가중치의 나이브 베이시안 방식에 비하여서도 4개의 데이터에 대하여 높은 성능을 보이고 있음을 알 수 있다. 이는 일반적인 나이브 베이시안의 성능도 어떠한 형태의 가중치를 고려한 나이브 베이시안과 비교할 때에 좋은 성능을 제공하는 경우도 많이 있음을 알 수 있었다.

7. 결 론

본 연구에서는 나이브 베이시안 학습에서 기존의 속성에 가중치를 부여하는 방식에서 한 걸음 더 나아가 속성의 값에 가중치를 부여하는 새로운 방식을 연구하였다. 이러한 속성 값의 가중치를 계산하기 위하여 Kullback-Leibler 함수를 이용하여 가중치를 계산하는 방식을 제안하였고 이러한 가중치들의 특성을 분석하였다. 제안된 알고리즘은 다수의 데이터를 이용하여 속성 가중치 방식과 비교하였고 대부분의 경우에 더 좋은 성능을 제공함을 알 수 있었다. 본 연구는 나이브 베이시안의 가중치 부여에서 새로운 연구의 방향을 제시한다고 볼 수 있다. 추후연구로는 더욱 정밀한 속성값 가중치 계산 방법을 개발하여서 나이브 베이시안에 적용할 수 있는 기술을 개발할 계획이다.

참 고 문 헌

[1] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. Machine Learning, 29(2-3), 1997.

[2] Pat Langley and Stephanie Sage. Induction of selective bayesian classifiers. In Proceedings of the Tenth Conference on Uncertainty in Articial Intelligence, pp.399-406, 1994.

[3] Claire Cardie and Nicholas Howe. Improving minority class prediction using case-specific feature weights. In Proceedings of the Fourteenth International Conference on Machine Learning, pp.57-65, 1997.

[4] Ron Kohavi and George H. John. Wrappers for feature subset selection. Articial Intelligence, vol.97, no.1-2, pp.273-324, 1997.

[5] C. A. Ratanamahatana and D. Gunopulos. Feature selection for the naive bayesian classier using decision trees. Applied Articial Intelligence, vol.17, no.5,6, pp.475-487, 2003.

[6] Thomas Gartner and Peter A. Flach. Wbcsvm: Weighted bayesian classification based on support vector machines. In ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, 2001.

[7] Mark Hall. A decision tree-based attribute weighting filter for naive bayes. Knowledge-Based Systems, vol.20, no.2, 2007. 13.

[8] Harry Zhang and Shengli Sheng. Learning weighted naive bayes with accurate ranking. In ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining, 2004.

[9] Dietrich Wettschereck, David W. Aha, and Takao Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Articial Intelligence Review, vol.11, pp.273-314, 1997.

[10] J. Ross Quinlan. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[11] Peter Clark and Robin Boswell. Rule induction with cn2: some recent improvements. In EWSL-91: Proceedings of the European working session on learning on Machine learning, pp.151-163, 1991.

[12] S. Kullback and R. A. Leibler. On information and suciency. The Annals of Mathematical Statistics, vol.22, no.1, pp.79-86, 1951.

[13] C. Merz, P. Murphy, and D. Aha. UCI repository of machine learning databases. 1997.

[14] U. Fayyad and K. Irani. *Multi-interval discretization of continuous-valued attributes for classification learning.* Proceedings of Thirteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, 1993.

이 창 환

정보과학회논문지 : 데이타베이스

제 37 권 제 4 호 참조