분류기 성능 향상을 위한 범주 속성 가상예제의 생성과 선별

(Generation and Selection of Nominal Virtual Examples for Improving the Classifier Performance)

이 유 정 † 강 병 호 †† 강 재 호 ††† 류 광 렬 ††††
(Yujung Lee) (Byoungho Kang) (Jaeho Kang) (Kwangryel Ryu)

요 약 본 논문에서는 베이지안 네트워크를 기반으로 생성하고 평가한 가상예제를 활용하여 범주 속성 데이타에 대한 분류 성능을 향상시키는 방안을 제안한다. 가상예제를 활용하는 종래의 연구들은 주로 수치 속성 데이타를 대상으로 하였고, 대상 도메인에 특화된 지식을 활용하여 특정 학습 알고리즘의 성능을 향상시키는 것을 목표로 하였다. 본 연구에서는 도메인에 특화된 지식을 활용하는 대신 주어진 훈련 집합을 기반으로 만든 베이지안 네트워크로부터 범주 속성 가상예제를 생성하고, 그 예제가 네트워크의 조건부 우도를 증가시키는데 기여할 경우 유용한 것으로 선별한다. 이러한 생성 및 선별과정을 반복하여 적절한 크기의 가상예제 집합을 수집하여 사용한다. 범주 속성 데이타를 대상으로 한 실험 결과, 여러 가지 학습 모델의 성능이 향상됨을 확인하였다.

키워드 : 기계학습, 분류, 베이지안 네트워크, 나이브 베이즈, 조건부 우도, 가상예제

Abstract This paper presents a method of using virtual examples to improve the classification accuracy for data with nominal attributes. Most of the previous researches on virtual examples focused on data with numeric attributes, and they used domain-specific knowledge to generate useful virtual examples for a particularly targeted learning algorithm. Instead of using domain-specific knowledge, our method samples virtual examples from a naïve Bayesian network constructed from the given training set. A sampled example is considered useful if it contributes to the increment of the network's conditional likelihood when added to the training set. A set of useful virtual examples can be collected by repeating this process of sampling followed by evaluation. Experiments have shown that the virtual examples collected this way can help various learning algorithms to derive classifiers of improved accuracy.

Key words : machine learning, classification, Bayesian network, naive Bayes, conditional likelihood, virtual example

1. 서 론

기계학습에서 분류란 주어진 훈련 예제들로 학습하여 생성한 분류기로 새로운 예제의 카테고리를 추정하는 것이다. 분류의 목적은 새로운 예제의 카테고리를 가능한 정확하게 예측하는 것이다. 분류 성능을 개선하기 위한 방안은 학습 알고리즘을 선택하거나 개선하는 방안과 훈련 집합을 변형하는 방안의 두 가지로 크게 나눌 수 있다. 첫 번째 방안인 학습 알고리즘을 선택하거나 개선하는 방안은 다시 다음과 같이 대략 세 가지로 구분된다. (1) 분류 문제에 따라 가장 효과적인 학습 알고리즘이 다를 수 있기 때문에 주어진 분류 문제에 적합한 학습 알고리즘을 선택함으로써 분류 성능을 개선한다. (2) 적용하고자 하는 학습 알고리즘의 학습 관련 인자를 분류 성능이 최대화되도록 적절히 조정한다. 이러한 학습 인자들의 예로 의사결정 나무 알고리즘(decision tree algorithm)에는 과부합을 방지하기 위하

· 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의해서 연구되었습니다.

† 학생회원 : 부산대학교 컴퓨터공학과
yjlee@pusan.ac.kr
†† 정 회 원 : 부산대학교 컴퓨터공학과
bhokang@pusan.ac.kr
††† 정 회 원 : 야후코리아 Search R&D센터
jhkang@pusan.ac.kr
†††† 종신회원 : 부산대학교 컴퓨터공학과 교수
krryu@pusan.ac.kr

논문접수 : 2006년 2월 16일
심사완료 : 2006년 10월 19일

여 가지치기(prunning)[1]를 수행할 지의 여부를 결정하는 확신도의 임계치가 있다. 또 다른 예로서 예제 기반 학습(instance-based learning)의 하나인 k-NN[2] 알고리즘에는 문제 예제와 가장 유사한 훈련 예제를 몇 개나 살펴볼 지를 설정하는 인자 k가 있다. (3) 이 이외에도 단일 분류기가 아니라 bagging[3], boosting[4], stacking[5]과 같이 여러 개의 분류기를 함께 활용하여 분류 성능을 개선하기도 한다.

분류 성능을 개선하기 위한 두 번째 방안은 훈련 집합을 변형하는 방안이다. 이러한 방안에는 분류에 도움이 되지 않는 잡음(noise) 예제를 제거하거나[6], 수치속성을 다루는 데에 비효율적인 학습 알고리즘을 사용하는 경우 사전에 수치속성을 이산화하는 방안이 있다[7]. 또한 기존 속성들을 조합하여 새로운 속성을 생성하여 사용하는 방안이 있다[8]. 이 외에도 전체 속성 중에서 분류에 효과적인 속성들을 선별하여 학습에 사용함으로써 분류 성능을 높이는 방안도 있다[9]. 본 논문에서는 가상예제를 생성하여 훈련 집합과 함께 학습에 활용함으로써 분류 성능을 향상시키는 방안을 제안한다. 본 논문 이전에도 가상예제를 분류 성능의 향상에 활용하는 방안은 다양하게 연구되어 왔다. 이들 연구에서는 도메인에 특화된 방법으로 가상예제를 생성하여 특정한 학습 알고리즘을 대상으로 분류 성능을 향상시켰고, 주로 수치속성 데이타를 대상으로 하였다. 그러나 본 제안 방안은 도메인에 독립적인 방법으로 가상예제를 생성하며, 이들 예제를 활용함으로써 여러 가지 분류기의 정확도를 향상시키는 것을 목표로 하고 있다.

가상예제란 주어진 훈련 집합에는 존재하지 않는 인공적으로 생성된 예제이다. 좋은 가상예제는 존재할 가능성이 높고 분류 성능 향상에 기여하는 것일 것이다. 본 논문에서는 존재할 가능성이 높은 가상예제를 생성하기 위해 주어진 훈련 집합을 기반으로 베이지안 네트워크를 만든 다음 이 네트워크에서 예제를 샘플링 한다. 좋은 가상예제의 두 번째 조건을 만족시키기 위해서는 생성된 가상예제가 훈련 집합의 분류 성능 향상에 효과적인지 평가가 필요하다. 이와 관련해서는 베이지안 네트워크의 조건부 우도(conditional likel:hood)가 증가하도록 네트워크를 변형시키면 분류의 목적인 클래스의 변별력이 향상된다는 기존 연구들이 있다[10,11]. 따라서 본 연구에서는 가상예제를 추가한 후에 수정된 베이지안 네트워크의 조건부 우도가 증가하면 이 가상예제가 분류 성능 향상에 기여한다고 판단하여 가상예제 집합에 추가한다. 이러한 과정을 반복하여 가상예제 집합을 구성한다.

본 연구는 한 학습 모델의 분류 성능을 향상시키는 예제들이 다른 학습 모델에서도 좋은 영향을 끼칠 것이라는 가설에서 출발하였다. 그러나 가상예제가 항상 분류 정확도 향상에 긍정적으로만 작용하는 것은 아니며, 또한 학습 알고리즘에 따라 가장 효과적인 가상예제의 규모가 다를 수 있다. 따라서 본 연구에서는 여러 가지 규모의 유용한 가상예제 집합을 미리 생성한 다음, 대상 학습 모델의 분류 성능 향상에 가장 유리한 규모의 집합을 정확도 추정기법과 통계적 기법을 통해 선별하였다. 선택된 가상예제 집합은 원래의 훈련 집합과 함께 학습에 사용된다. 본 논문에서 제안한 방안이 나이브 베이즈(naive Bayes)[12], 결정 나무(decision tree)[1], 서포트 벡터 머신(support vector machine)[13], 1-NN (nearest neighbours)[2]와 같은 여러 학습 알고리즘의 성능 개선에 효과가 있음을 실험을 통하여 확인하였다.

본 논문의 2장에서는 가상예제와 관련된 기존 연구들을 정리하고, 3장에서는 가상예제들을 생성하고 평가하는 방안을 자세히 설명한다. 4장에서는 생성한 여러 개의 가상예제 집합 중에서 분류 성능 향상에 가장 효과적인 가상예제 집합을 선택하는 방안을 기술한다. 5장에서는 UCI 데이타를 대상으로 한 실험결과를 정리하여 분석하고, 6장에서는 결론과 향후 연구에 대하여 기술한다.

2. 관련 연구

가상예제를 이용하여 신경망의 일반화 능력을 향상시키는 방안에 관한 연구가 있었다[14-16]. 이 방안은 예제가 분포된 공간의 밀도를 추정한 후, 예제가 희박한 영역에 가상예제를 생성하여 학습에 활용하였다. 가상예제의 클래스는 실제 훈련 예제로 학습하여 생성한 신경망으로 결정하였다. k-NN 학습 알고리즘을 대상으로 가상예제를 생성하는 여러 가지 방안을 소개한 연구가 있었다[17]. 동일한 클래스에 속하는 두 개의 훈련 예제들을 임의로 선택하여 서로의 속성값을 평균하여 가상예제를 생성하거나 임의로 선택한 한 개의 훈련 예제를 약간의 노이즈를 추가하여 가상예제의 속성값을 결정하고 생성에 사용된 훈련 예제의 클래스를 가상예제의 클래스 설정에 사용하였다. 인공 분류 문제를 포함하여 여러 분류 문제에서 k-NN 분류기의 성능 향상에 효과가 있음을 실험적으로 보였다. 나이브 베이즈 학습 알고리즘의 분류 성능 개선을 위하여 가상예제를 활용한 연구가 있었다[18]. 이 연구에서는 무작위로 가상예제를 생성하고 leave-one-out 검증을 이용하여 초기 훈련 예제의 정답 클래스들에 대한 추정 확률을 높이는 가상예제들을 선별하여 훈련 집합과 함께 활용하였다. 그 결과 여러 분류 문제에서 나이브 베이즈 학습 알고리즘의 성능 향상의 가능성을 보였다. 이 이외에도 서포트 벡터 머신에서도 가상예제와 비슷한 개념인 가상 서포트 벡터를 생성하여 숫자 인식문제에 적용한 연구가 있었다[19]. 서포트 벡터 머신 학습을 통하여 서포트 벡터들를

찾고, 찾은 서포트 벡터에 변화를 주어 가상 서포트 벡터를 생성하여 분류 정확도를 개선하였다.

적용하고자 하는 분류 문제의 특성을 이용하여 가상 예제를 생성하여 활용한 연구들이 있었다. 얼굴 인식 성능 향상을 위하여 하이브리드 분류기에 가상예제를 활용하는 방안이 연구되었다[20]. 이 연구에서는 실제 얼굴사진을 좌우 반전하여 가상의 사진을 생성하거나 동일한 사람을 촬영한 두 개의 실제 사진을 이용하여 그 속성값들의 평균하여 가상의 사진을 생성하였다. 가상의 사진에 대한 카테고리는 생성에 사용한 실제 사진들과 동일하게 결정하였다. 이러한 방안으로 생성한 가상 사진 즉 가상예제는 얼굴 인식 성능을 향상시킬 수 있었다. 최근에는 자동차 번호판 인식 문제에서 수집된 훈련 예제들의 속성값들을 평균하여 가상예제를 생성하는 방안이 제안되었고 그 효용성이 확인되었다[21]. 그리고 문서분류에 있어서 주어진 문서들을 결합해서 생성된 문서를 이용하여 성능을 향상시킨 연구가 있었다[22].

기존 연구들에서 가상예제가 여러 학습 알고리즘에서 분류 성능 향상에 도움이 되는 것을 확인할 수 있었다. 하지만 기존 연구들은 주로 수치 속성으로만 표현된 분류 문제를 대상으로 하고 있고 가상예제 생성 방안이나 평가 방안이 특정 알고리즘 또는 특정 분류 문제에 특화된 연구들이었다. 본 연구에서는 이전 연구와는 다르게 가상예제가 범주 속성만으로 표현되는 분류 문제들과 여러 학습 알고리즘의 분류 정확도에 효과가 있음을 실험적으로 밝히고자 한다.

3. 가상예제 집합 생성

본 제안 방안은 크게 두 단계로 나누어진다. 첫 번째는 가상예제 집합을 생성하는 단계이이고 두 번째는 첫 번째 단계를 반복 수행하여 생성된 여러 후보 가상예제 집합 중에서 적용하고자 하는 학습 알고리즘의 분류 성능 향상에 가장 효과적인 것을 선별하는 단계이다. 본 장에서는 먼저 분류 성능 향상에 도움이 되는 가상예제들을 생성하여 가상예제 집합을 구성하는 방안을 자세히 설명한다.

3.1 베이지안 네트워크

가상예제를 도메인과 독립적으로 생성하는 가장 간단한 방안으로는 가상예제의 속성 값을 임의로 생성하는 것을 생각할 수 있다. 그러나 임의로 생성한 가상예제들 중에서는 네크워크의 조건부 우도를 향상시키는데 기여하는 경우가 너무 희박하게 발견되어 임의 생성방안은 실효성이 거의 없음을 실험적으로 확인하였다. 따라서 도메인에 존재할 법한 가상예제를 생성하는 방안으로 주어진 훈련예제들을 기반으로 베이지안 네트워크를 만들어 가상예제를 생성하는 방안을 생각하게 되었다.

베이지안 네트워크[11]는 노드(node)들의 집합 V와 아크(arc)들의 집합 A, 그리고 각 노드별 조건 확률표들의 집합 Θ으로 표현되는 방향성 비순환 그래프(directed acyclic graph) $B=\{V, A, \Theta\}$이다. 베이지안 네트워크에서 노드들의 집합 $V=\{x_1,...,x_v\}$는 랜덤 변수(random variable)들을 나타내고 방향성을 가진 아크들은 노드간의 의존성(dependency)를 나타낸다. 베이지안 네트워크상에서 노드 x_i로 아크가 그어져 있는 노드를 x_i의 부모 노드(parents node)라고 하고 x_i의 부모 노드들의 집합을 π_i라 정의한다. x_i의 부모 노드에 해당되는 변수들의 값이 결정된 경우 자손에 해당하지 않는(non-descendent) 노드들과 x_i는 독립이다. 그림 1은 베이지안 네트워크의 한 예를 보이고 있다. 그림 1에서 d_{ij}는 x_i가 가질 수 있는 값 중에서 j번째 값을 의미한다. p_{ijk}는 x_i의 부모 노드가 j번째 값을 가질 때 x_i가 k번째 값을 가질 확률을 의미한다.

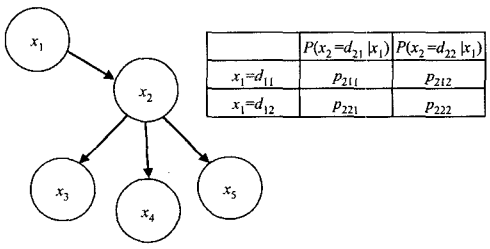


그림 1 베이지안 네트워크의 예

훈련 집합을 $D=\{X_1,...,X_d,...,X_n\}$라고 하고 임의의 이벤트를 $X_d=(x_{d,1},...,x_{d,v})$이라 하면 X_d의 존재 확률은 수식 (1)처럼 정의된다.

$$P_B(X_d)=\prod_{i=1}^{v} P(x_{d,i}|\pi_{d,i})) \quad (1)$$

베이지안 네트워크를 생성하기 위해서는 두 가지를 정의해야 한다. 주어진 훈련 집합의 존재 확률이 가장 높은 베이지안 네트워크 구조와 각 노드들의 조건 확률표의 원소들의 값, 즉 p_{ijk}를 결정해야 한다. 훈련 집합의 존재 확률은 우도(likelihood)라고 하고 수식 (2)와 같이 정의된다.

$$L(B|D)=\prod_{d=1}^{n} P_B(X_d)=\prod_{d=1}^{n}\prod_{i=1}^{v} P_B(x_{d,i}|\pi_{d,i}) \quad (2)$$

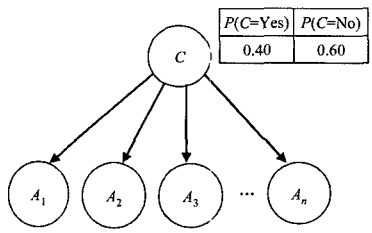
수식 (2)와 같은 우도가 최대가 되도록 베이지안 네트워크를 구성하는 경우 이를 생산적 모델(generative model)이라고 한다. 생산적 모델을 생성하고자 할 경우 일단 네트워크 구조만 정해지면 각 조건 확률표의 원소 p_{ijk}는 관측 횟수 추정(observed frequency estimates)

($\widehat{p_{ijk}} = n_{ijk}/n_{ij}$)을 이용하여 간단히 설정할 수 있다. 여기에서 n_{ijk}는 훈련 집합에서 x_i의 부모 노드가 j번째 값을 가질 때 x_i가 k번째 값을 가지는 훈련 예제들의 수를 의미하고, n_{ij}는 x_i의 부모 노드가 j번째 값을 가지는 훈련 예제들의 수를 의미한다. 훈련 집합의 존재 확률이 최대화되도록 구성한 생산적 베이지안 네트워크에서 가상예제들을 생성한다면 존재 확률이 높은 좋은 가상예제들을 얻을 수 있을 것이다.

3.2 후보 가상예제 생성

본 논문에서 베이지안 네트워크 구조로는 여러 분류 문제에 널리 사용하는 나이브 베이즈 모델[12]을 사용하였다. 나이브 베이즈는 그림 2와 같이 각 속성 노드는 클래스 노드만 부모 노드로 가지고, 임의의 노드는 클래스 노드 이외의 다른 노드들과 독립을 가정한다. 나이브 베이즈는 구조가 간단하지만 분류 성능이 우수하고 계산 비용이 적어 여러 분류 문제에 많이 적용되어 왔다. 나이브 베이즈 모델에서 가상예제를 생성하는 방안은 다음과 같다. 해당 노드의 조건 확률표를 이용하여 최상위 노드인 클래스 노드의 값을 결정한 후 각 속성의 값을 결정한다.

그림 2에서 가상예제 생성의 예를 들자면, 가상예제의 클래스가 Yes로 결정한 후에 3번째 속성 A_3의 값을 결정하고자 할 때, 조건 확률표에서 A_3의 속성값은 0.35의 확률로 Sunny가 결정되고 0.65의 확률로 Windy가 결정된다. 위와 같은 방법으로 가상예제의 클래스와 모든 속성값이 결정되면 하나의 후보 가상예제가 생성된다. 여기에서 후보 가상예제란 나이브 베이즈 모델을 통하여 모든 속성값과 클래스를 결정하여 생성하였지만 분류 성능 향상에 기여하는지 평가가 필요한 가상예제를 말한다.



$P(C$=Yes)	$P(C$=No)
0.40	0.60

	$P(A_3$=Sunny|C)	$P(A_3$=Windy|C)
C=Yes	0.35	0.65
C=No	0.90	0.10

그림 2 나이브 베이즈 모델의 예

3.3 후보 가상예제 평가

본 절에서는 생성된 후보 가상예제가 분류 성능 향상에 기여하는지 평가하는 방안을 설명한다. 본 논문에서 사용한 가상예제의 평가 척도는 나이브 베이즈로 구현된 베이지안 네트워크의 조건부 우도(conditional likelihood)이다. 3.1절에서 베이지안 네트워크를 구성하는 방안으로 우도를 최대화하는 생산적 모델을 설명하였다. 이러한 생산적 모델은 훈련 예제들의 존재 확률을 최대화하도록 베이지안 네트워크를 구성한다. 하지만 분류의 목적은 임의의 예제 X_d의 속성값들이 주어졌을 때 X_d의 클래스를 가능한 정확하게 추정하는 것이다. 베이지안 네트워크에서 X_d의 속성값의 벡터 $(x_{d,1},\cdots,x_{d,v-1})$가 주어지면 $y_d(=x_{d,v})$는 $P(y_d|x_{d,1},\cdots,x_{d,v-1})$이 가장 큰 클래스로 추정한다. 최근에는 베이지안 네트워크의 분류 성능을 향상시키기 위한 연구 중에서 수식 (3)과 같은 조건부 우도를 최대화하는 판별적 모델(discriminative model)이 연구되었다[10,11].

$$CL(B|D) = \prod_{d=1}^{n} P_B(y_d|x_{d,1},\cdots,x_{d,v-1}) \quad (3)$$

분류의 용도로 베이지안 네트워크를 사용하고자 할 경우에는 수식 (2)의 우도보다 수식 (3)의 조건부 우도를 최대화하는 하는 것이 클래스 변별력을 높여 분류에 더 유리하다. 하지만 우도와 달리 조건부 우도를 최대화하도록 각 노드의 조건 확률표를 설정할 수 있는 간단한 방법이 존재하지 않기 때문에 휴리스틱 또는 탐색을 적용한 연구들이 있다[10,11]. 이러한 연구들을 바탕으로 본 논문에서는 생성된 가상예제를 추가하여 업데이트한 베이지안 네트워크의 조건부 우도가 가상예제를 추가하기 전보다 증가한다면 해당 가상예제가 클래스 변별력 향상에 효과가 있을 것으로 판단된다. 본 연구에서는 조건부 우도를 계산할 때 초기 훈련 집합만을 이용해서 계산하였다.

3.4 가상예제 집합 생성 알고리즘

그림 3은 지금까지 설명한 가상예제 집합 생성 과정을 알고리즘으로 정리한 것이다. 함수 GenerateVirtualExampleSet(T,n)은 입력으로 훈련 집합 T와 생성할 가상예제들의 개수 n을 받아 n개의 가상예제를 가진 집합 V를 생성하여 출력한다. 구체적으로 알고리즘 내에 존재하는 변수들의 정의는 다음과 같다. v는 한 개의 후보 가상예제이고 T'는 현재까지 생성해서 선택된 가상예제들과 초기 훈련 예제들의 합집합이다. $B_{T'}$는 T'로 생성한 베이지안 네트워크이며 $CL_{T'}$는 $B_{T'}$의 조건부 우도이다. BuildnaiveBayes(T')는 T'를 이용하여 나이브 베이즈를 기반으로 하는 베이지안 네트워크 $B_{T'}$를 생성하는 함수이고 CalculateCL($B_{T'}$,T)는 $B_{T'}$의 조건부 우도를 계산하는 함수이다.

4. 가상예제 집합 선별

3장에 소개한 가상예제 집합 생성을 반복 수행하면

```
procedure GenerateVirtualExampleSet
input : T - training set, n - the number of virtual example to be generated
output : V - virtual example set
begin
  V ← ∅
  T' ← T
  B_T ← BuildnaiveBayes(T'), CL_T ← CalculateCL(B_T,T)
  while ( |V| < n )
    v ← RandomSampling(B_T)
    B_{T'∪{v}} ← BuildnaiveBayes(T'∪{v})
    CL_{T'∪{v}} ← CalculateCL(B_{T'∪{v}},T)
    if ( CL_{T'∪{v}} > CL_T )
      T' ← T'∪{v}
      B_T ← B_{T'∪{v}}
      CL_T ← CL_{T'∪{v}}
      V ← V∪{v}
    end if
  end while
  return V
end
```

그림 3 가상예제 집합 생성 알고리즘

여러 개의 가상예제 집합들을 얻을 수 있다. 여러 개의 가상예제 집합을 생성하는 이유는 베이지안 네트워크에서 가상예제들이 랜덤 샘플링을 통해 생성되므로 각기 다른 가상예제 집합이 나오고 또한 얼마만큼의 가상예제를 사용하는 것이 좋은지 알 수 없기 때문이다. 각기 다른 가상예제 집합은 성능도 다르기 때문에 적용하고자 하는 학습 알고리즘에 가장 적합한 가상예제 집합을 선택하는 과정을 본 장에서 자세히 소개한다.

4.1 정확도 추정을 통한 가상예제 집합 선별

기계학습에서 분류기의 정확도를 추정하는 방안들로 분할 교차 검증 방법(cross validation), leave-one-out, hold-out, bootstrap등이 있다. 본 논문에서는 정확도 추정에 신뢰도가 높아 널리 사용되는 10-10 분할 교차 검증을 적용한다. 10-10 분할 교차 검증이란 10 분할 교차 검증(ten-fold cross validation)을 10회 반복하는 방안이다. 하지만 본 논문에서 사용하는 방법은 일반적인 10 분할 교차 검증과 수행 방법에 약간의 차이점이 있다. 가상예제 집합의 유용성을 검증하기 위하여 학습을 위한 데이타에만 가상예제 집합을 사용하고, 평가를 위한 데이타에는 가상예제 집합을 사용하지 않는다.

각각의 가상예제 집합을 추가하기 전후에 생성한 분류기들의 성능 비교 결과가 통계적으로 의미를 가지는지 확인하기 위하여 t-검증을 사용한다. t-검증을 통과한 가상예제 집합들 중에서 추정 정확도가 가장 높은 가상예제 집합을 최종적으로 선택한다. 만약 t-검증을 통과한 가상예제 집합이 없다면 가상예제 집합을 선택하지 않고 초기 훈련 집합으로 분류기를 생성한다.

4.2 가상예제 집합 선별 알고리즘

그림 4는 가상예제 집합 선별 알고리즘이다. 입력으로는 가상예제 집합들의 집합 $\boldsymbol{V}$, 훈련 집합 T, 적용할 학습 알고리즘 L이 주어지고 최종적으로 선택한 가상예제 집합 V_B가 출력된다. 그림 4에서 V_0는 공집합으로 가상예제 집합을 사용하지 않는 경우이고 c는 t-검증의 유의수준이다. EstimateAccuracy(T, V_k, L)은 가상예제 집합 V_k를 훈련 집합과 함께 활용하여 생성한 분류기의 정확도를 추정하는 함수이다. PairedTtest(*listofaccuracy*$_0$, *listofaccuracy*$_k$, c)는 EstimateAccuracy(T, V_k, L)에서 출력으로 얻은 정확도들을 입력으로 받아 유의 수준 c로 t-검증를 수행한다. 생성된 모든 가상예제 집합이 분류 성능 향상에 도움이 되지 않다고 판단된 경우에는 공집합인 V_0가 출력된다.

그림 5에는 본 제안 방안의 전체 과정을 알고리즘으로 기술하였다. 함수 LearningwithVirtualExamples의 입력으로는 훈련 집합 T, 사용할 학습 알고리즘 L, 신뢰수준 c, 생성할 가상예제 집합의 개수와 크기목록인 *listofpercentage*가 들어가고 출력으로는 하나의 분류기가 나오게 된다.

가상예제 집합을 생성하고 선별하는 데 걸리는 시간 복잡도는 생성과 선별과정으로 나누어서 계산할 수 있다. 훈련예제 개수가 n, 속성 개수가 m, 생성할 가상예제 수가 v, 클래스 개수가 c, 속성들 중 속성값 최대 개수가 a일 때 나이브 베이즈 네트워크를 생성하는 데 걸리는 시간 복잡도는 $O(mca)$ 이다. 그리고 하나의 가상예제를 생성하는 시간 복잡도는 $O(ma)$이고 가상예제로

```
procedure SelectVirtualExampleSet
input : V = {V_0, V_1, ..., V_k, ..., V_m}, T - training set, L - learning algorithm
        c - significance level
output : V_B - the best virtual example set
begin
  Vc ← ∅
  listofaccuracy_0 ← EstimateAccuracy(T, V_0, L)
  foreach V_k in V (k ≠ 0) do
    listofaccuracy_k ← EstimateAccuracy(T, V_k, L)
    if ( Average(listofaccuracy_0) < Average(listofaccuracy_k) )
      if ( PairedTtest(listofaccuracy_0, listofaccuracy_k, c) )
        Vc ← Vc∪V_k
      end if
    end if
  end foreach
  if ( Vc ≠ ∅ )
    V_B ← SelectBest(Vc)
  else
    V_B ← V_0
  end if
  return V_B
end
```

그림 4 가상예제 집합 선별 알고리즘

```
procedure LearningwithVirtualExamples
input : T - training set, L - learning algorithm, c - confidence level
        listofpercentage - the list of size for virtual example sets
output : h_L - a classifier
begin
  V ← ∅
  foreach x in listofpercentage do
    V_i ← GenerateVirtualExampleSet(T,|T|×x)
    V ← V∪{V_i}
  end foreach
  V ← SelectVirtualExampleSet(T,L,V,c)
  h_L ← Learning(T,L,V)
  return h_L
end
```

그림 5 가상예제 추가방안 전체 알고리즘

네트워크를 업테이트하여 조건부 우도를 계산하는 시간 복잡도는 $O(m+mca)$이다. 생성된 가상예제 중 선별과정을 통과하는 가상예제의 비율을 p라고 하면 가상예제 집합 생성과정의 전체 시간 복잡도는 $O(v(1/p)mca)$가 된다.

가상예제 집합을 선택하는 과정에서는 학습 횟수를 통해 시간 복잡도를 추정할 수 있다. 평가받을 가상예제 집합 개수가 s이고 학습 알고리즘 A로 학습하는 데 걸리는 시간 복잡도를 l_A 라고 하면, s 개의 가상예제 집합 중 가장 좋을 것으로 추정되는 집합을 10 분할 교차 검증 10회 반복하여 선택하는 데에는 $100sl_A$ 번의 학습이 필요하다.

5. 실험 결과 및 분석

본 제안 방안이 효과가 있는지 확인하기 위하여 대상 분류 문제로 UCI 범주 속성 데이타 11개를 실험에 사용하였다[23]. 표 1은 실험 데이타들의 특성을 나열하고 있다.

본 제안 방안이 어떤 학습 알고리즘에 효과가 있는지 알아보기 위하여 다음과 같이 나이브 베이즈(NB), 의사결정 나무 학습(C4.5)[1], 서포트 벡터 머신 종류의 하나인 SMO[13], 예제 기반 학습의 하나인 1-NN(nearest neighbours)[2] 학습 알고리즘을 적용하였다. 분류기는 Weka 데이타 마이닝(data mining) 소프트웨어를 이용하여 생성하였다[24][25]. 가상예제 집합은 주어진 훈련 집합의 크기를 대비하여서 50%, 100%, 150%, 200%, 250%, 300%, 350%, 400%, 450%, 500%로 총 10개를 생성하였다. 본 제안 방안이 효과가 있는지 검증하기 위하여 전체 학습 과정에 대하여 10분할 교차 검증[1]을 10회 수행하여 나온 결과들의 평균하였다.

표 1 실험 데이타 특성

데이타명	예제 수	범주 속성 수	클래스 수
audiology	226	69	24
breast-c	286	9	2
kr-vs-kp	3196	36	2
monks-1	556	6	2
monks-2	601	6	2
monks-3	554	6	2
primary	339	17	25
soybean	683	35	19
splice	3190	62	3
vote	435	16	2
zoo	101	17	7

그림 6의 네 개의 그래프는 각각 NB, C4.5, SMO, 1-NN 학습 알고리즘에 대해서 11개의 데이타를 실험한 결과이다. 가로축은 주어진 훈련 집합으로만 학습하여 생성한 분류기의 추정 정확도이고 세로축은 본 제안 방안으로 생성한 분류기의 추정 정확도이다. 각 그래프의 한 점은 하나의 데이타에 대한 실험 결과를 나타낸다. 그래프 상에 $y=x$ 직선을 기준으로 직선 위쪽에 있는 점들은 본 제안 방안을 적용하였을 때 더 우수한 분류 성능을 보인 경우이다. 가상예제를 활용하는 방안은 NB뿐만 아니라 C4.5, 1-NN 학습 알고리즘의 성능 향상에 대체적으로 기여함을 알 수 있다.

표 2는 그림 6을 표로 나타낸 것이다. 표 2에서 NB는 나이브 베이즈 학습 알고리즘이고 NB_V는 나이브 베이즈 학습 알고리즘에 본 논문에서 제안한 방안을 적용한 것이다. 그 이외의 알고리즘들도 모두 같은 표기 형식을 사용하였다. 표에는 각 분류기의 정확도와 표준편차를 기록하였다. 위쪽 화살표는 가상예제 집합을 활용하여 생성한 분류기가 95%의 신뢰도로 더 우수한 성능을 보인 경우이다. 아래쪽 화살표는 반대로 분명한 성능저하가 일어난 경우이다. 그리고 표 2의 가장 아랫줄에는 알고리즘별 평균 정확도를 계산하여 놓았다.

표 2를 살펴보면, audiology의 경우 주어진 훈련 집합으로만 학습했을 때 NB가 다른 3개의 알고리즘보다 성능이 떨어졌다. 하지만 본 제안 방안을 적용하고 난 후에는 가상예제를 추가하지 않은 다른 학습 알고리즘보다 성능이 우수한 것을 볼 수 있다. 그리고 대부분의 데이타에서 NB의 성능이 향상된 경우에는 1-NN과 C4.5의 성능도 함께 향상된 것을 볼 수 있다. 평균 향상율을 보면 NB가 1.79%로 가장 많이 올랐고 그 다음으로 1-NN이 1.52%이고, C4.5가 1.24%이다. SMO의 평균향상율은 0.12%로 다른 학습 알고리즘에 비해 정확도

1) 4장에서 설명한 10 분할 교차 검증과는 다르다. 4장에서 소개한 10 분할 교차 검증은 본 제안 방안의 알고리즘 내부에서 유용한 가상예제 집합을 선택하기 위한 장치였다. 5장에서 소개하는 10분할 교차 검증은 본 제안 방안 전체를 검증하기 위한 장치이다.

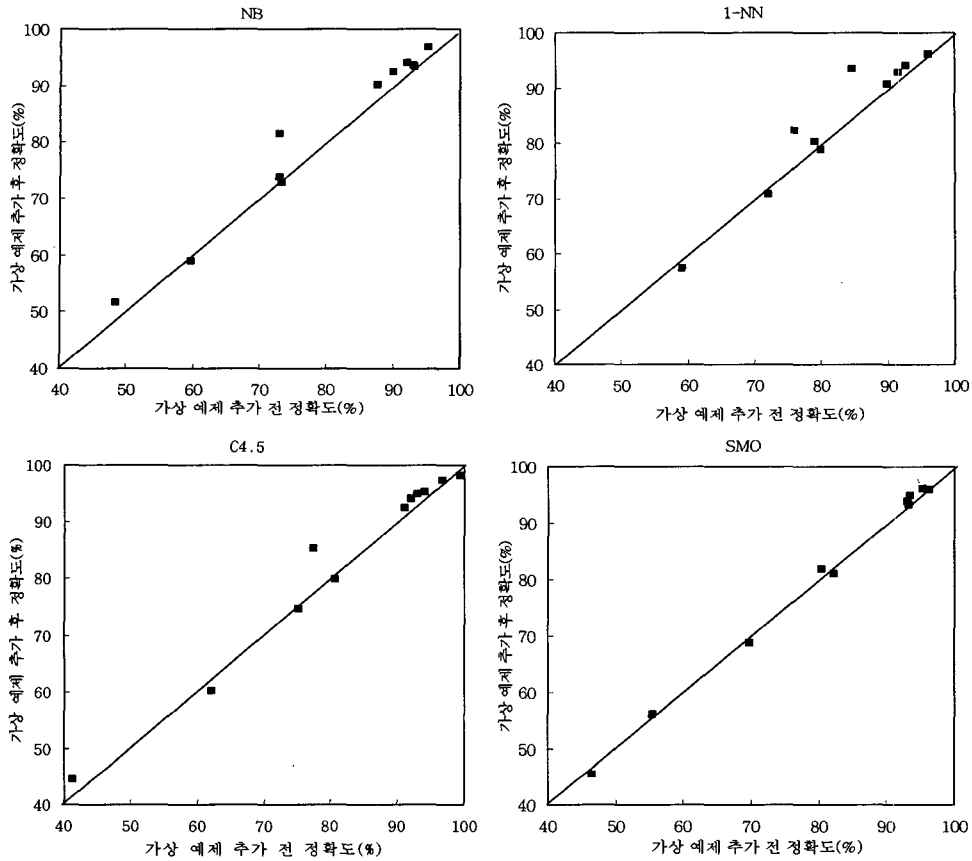


그림 6 알고리즘별 가상예제 추가 전후의 분류기 성능 비교

표 2 가상예제 집합 추가방안을 적용하였을 때 분류 정확도 비교표

정확도(%) 표준편차	NB	NB_V	1-NN	1-NN_V	C4.5	C4.5_V	SMO	SMO_V
audiology	72.9	↑ 81.4	75.2	↑ 80.2	77.5	↑ 85.4	81.5	↑ 84.8
breast-c	71.6	72.3	75.5	75.2	75.2	74.5	69.9	↓ 68.8
monks-1	77.4	76.9	72.8	73.3	82.3	↓ 79.9	83.9	82.9
monks-2	56.7	58.3	55.2	↓ 53.3	56.2	↓ 54.1	58.6	57.2
monks-3	93.3	93.3	77.1	↑ 83.3	93.4	93.6	93.3	93.3
kr-vs-kp	87.7	↑ 93.1	90.0	90.8	99.2	98.6	95.4	↑ 96.9
primary	50.1	↑ 52.7	33.6	↑ 39.9	43.5	↑ 44.6	46.9	↓ 45.4
soybean	92.2	↑ 93.8	89.9	↑ 92.9	91.5	↑ 94.1	93.2	↑ 94.9
splice	95.3	↑ 96.7	75.9	↑ 82.4	94.1	↑ 95.4	93.4	↑ 94.9
vote	90.1	↑ 92.3	92.6	↑ 93.6	96.3	97.4	96.1	95.9
zoo	93.0	↑ 93.5	96.0	96.0	92.1	↑ 95.0	96.0	96.0
평균 정확도 (평균 향상율)	80.0	82.2 (2.2)	75.8	77.8 (2.0)	81.8	82.9 (1.1)	82.6	82.8 (0.2)

향상 정도가 낮았다.

나이브 베이즈와 표면적으로 관계가 없어 보이는 1-NN, C4.5에서도 분류 정확도가 향상되는 것은 주목할 만한 현상이다. 일반적으로 분류의 목적은 예제들의 클래스를 정확하게 분별하는 것이고 이것은 모든 학습 알고리즘들의 공통된 목표라고 할 수 있다. 본 논문에서 제안한

생성한 가상예제들이 기존 훈련 예제들의 클래스 변별력을 높여주는 효과를 가져 NB 이외의 다른 학습 알고리즘의 성능 향상에도 기여하는 것으로 보인다.

표 3은 표 2를 정리한 것이다. win은 본 제안 방안이 t-검증을 이용하여 분명한 향상 정도를 보인 데이타의 개수이고 loss는 분명한 성능 저하가 일어난 데이타의 개수이다. 본 제안 방안을 적용하여 성능이 개선되었거나 저하되었지만 t-검증에서 통과하지 못한 데이타들의 개수를 draw에 표기하였다. NB, 1-NN, C4.5의 성능은 대체로 개선되었고 SMO는 다른 학습 알고리즘에 비하여 크게 개선되지 않았다. 이상의 실험을 통하여 본 제안 방안이 실용성이 있음을 확인하였다.

표 4는 학습 알고리즘이나 분류 문제 별로 가상예제 집합을 사용한 비율이 어떻게 다른지를 보이기 위해 100번의 실험 중 가상예제를 선택한 경우의 횟수를 정리한 것이다. 이 표를 보면 대체적으로 성능 향상이 많은 경우에 가상예제 선택 비율도 높은 것을 알 수 있다. 예를 들어, monks-3 데이타에 대해 뚜렷한 성능 향상을 보이지 못한 NB와 SMO의 경우 (표 2 참조) 가상예제 집합의 선택비율이 낮은 반면 가상예제를 사용하여 큰 성능향상을 보인 1-NN과 C4.5의 경우에는 가상예제 집합의 선택 비율도 높다.

표 3 실험 결과 요약

	NB	1-NN	C4.5	SMO
win	7	6	6	4
draw	4	4	3	5
loss	0	1	2	2

표 4 가상예제 집합 사용 횟수

Data set	NB	1-NN	C4.5	SMO
audiology	93	96	91	82
breast-c	65	45	63	45
kr-vs-kp	67	75	46	57
monks-1	21	24	25	21
monks-2	23	12	42	33
monks-3	6	76	78	8
primary	75	85	57	56
soybean	65	75	65	56
splice	68	63	71	74
vote	65	43	46	43
zoo	21	4	35	2

표 5는 가상예제 집합을 선택하는 과정에서 시행하는 t-test의 신뢰도 구간을 0%, 25%, 50%, 60%, 70%, 80%, 90%, 95%, 99%로 달리해 가며 실험한 결과를 보인 것이다. win은 실험에서 사용한 네 가지의 알고리즘의 11개 데이타에 대한 분류 성능이 분명히 향상된 경우이고, loss는 분류성능이 분명히 저하된 경우이다. 신뢰도가 90%일 때가 loss가 가장 적고 win 이 가장 많기는 하지만, 대체로 보아 t-test의 신뢰도에 크게 민감하지 않을 뿐 아니라 t-test를 하지 않더라도 (표에 0%로 표시) 큰 차이는 없음을 볼 수 있다. 이는 가상예제 생성과 평가단계를 통해 양질의 가상예제들이 확보되었음을 의미하는 것이다.

6. 결론 및 향후 연구

본 논문에서는 분류 성능을 향상시키기 위하여 가상예제를 생성하고 평가하여 훈련 집합과 함께 학습에 활용하는 방안을 소개하였다. 훈련 예제들을 이용하여 구성한 베이지안 네트워크로 존재 확률이 높은 가상예제를 생성하고 조건부 우도로 가상예제가 분류에 유용한지 평가하였다. 이렇게 분류 성능 향상에 도움이 되는 가상예제만을 선택하여 가상예제 집합을 생성하고, 적용하고자 하는 학습 알고리즘에 대해 어떤 가상예제 집합이 적합한지 통계적 검증을 통해 선별하여 학습에 활용하였다. 실험 결과 가상예제가 비록 나이브 베이즈에서 생성되었지만 여러 가지 다른 학습 알고리즘이 보다 정확한 분류기를 유도해 내는데 도움을 줄 수 있음을 확인하였다.

실세계에는 범주속성으로 이루어진 분류 문제 이외에도 수치 속성으로 이루어진 많은 분류 문제들이 존재한다. 앞으로 본 논문에서 제안한 방안을 확장하여 수치 속성을 포함한 분류 문제들에 적용하는 연구가 필요할 것으로 본다. 또한 본 연구에서는 나이브 베이즈를 사용하여 가상예제를 생성하였지만 보다 효과적인 구조의 베이지안 네트워크를 활용한다면 더 유용한 가상예제를 생성할 수 있을 것으로 기대한다.

참 고 문 헌

[1] Quinlan, J. R., C4.5 : Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.

[2] Aha, D. and Kibler, D., "Instance-based Learning

표 5 신뢰도 구간에 따른 실험

Confidence level	0%	25%	50%	60%	70%	80%	90%	95%	99%
Win	24	25	24	26	26	28	29	28	28
Draw	23	23	24	22	22	21	21	21	22
Loss	8	7	7	7	7	3	5	6	5

Algorithms," Machine Learning, Vol.6, pp. 37-66, 1991.
[3] Breiman, L., "Stacked Regression," Machine Learning, Vol.24, No.2, pp. 123-140, 1996.
[4] Freund, Y. and Schapire, R. E., "Experiments with a New Boosting Algorithm," Proc. of the 13th International Conference on Machine Learning, pp. 148-156, 1996.
[5] Wolpert, D. H., "Stacked Generalization," Neural Networks, Vol.5, pp. 241-259, 1992.
[6] Aha, D. W., "Tolerating Noisy, Irrelevant, and Novel Attributes in Instance-based Learning Algorithms," International Journal of Man-Machine Studies, Vol.36, No.2, pp. 267-287, 1992.
[7] Kohavi, R. and Sahami, M., "Error-based and Entropy-based Discretization of Continuous Features," Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 114-119, 1996.
[8] Pazzani, M., "Constructive induction of Cartesian product attributes," Information, Statistics and Induction in Science, pp. 66-77, 1996.
[9] Almuallim, H. and Dietterich, T. G., "Learning With Many Irrelevant Features," Proc. of the 9th National Conference on Artificial Intelligence, pp. 547-552, 1991.
[10] Greiner, R. and Zhou, W., "Structural Extension to Logistic Regression: Discriminative parameter learning of belief net classifiers," Proc. of the 18th National Conference on Artificial Intelligence, pp. 167-173, 2002.
[11] Grossman, D. and Domingos, P., "Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood," Proc. of the 21th International Conference on Machine Learning, pp. 361-368, 2004.
[12] John, G. and Langley, P., "Estimating Continuous Distributions in Bayesian Classifiers," Proc. of the 11th Conference on Uncertainty in Artificial Intelligence, pp. 338-345, 1995.
[13] Scholkopf, B., Burges, C. J. C. and Smola, A. J., Advance in Kernel Methods - Support Vector Learning, MIT Press, 1998.
[14] Sietsma, J. and Dow, R. J. F., "Creating Artificial Neural Networks that Generalize. Neural Networks," IEEE transactions on Neural Networks, Vol.4, pp. 67-79, 1991.
[15] Cho, S. and Cha, K., "Evolution of Neural Network Training Set through Addition of Virtual samples," Proc. of the 1996 IEEE International Conference on Evolutionary Computation, pp. 685-688, 1996.
[16] Cho, S., Jang, M. and Chang, S., "Virtual Sample Generation using a Population of Networks," Neural Processing Letters, Vol.5, No.2, pp. 83-89, 1997.
[17] 김종성, "분류 성능 향상을 위한 가상예제 생성 방안", 부산대학교 석사학위논문, 2004.
[18] 이유정, 강병호, 강재호, 류광렬, "가상예제를 이용한 naive Bayes 분류기 성능 향상", 한국정보과학회 제32회 추계학술발표회 논문집, Vol.32, No.2, pp. 655-657, 2005.
[19] Burges, C. and Scholkopf, B., "Improving the Accuracy and Speed of Support Vector Machines," Advances in Neural Information Processing System, Vol.9, No.7, 1997.
[20] Ryu, Y. S. and Oh, S. Y., "SIMPLE Hybrid Classifier for Face Recognition with Adaptively Generated Virtual Data," Pattern Recognition Letters, 2002.
[21] 김종성, 박태진, 강재호, 백남철, 강원희, 이상협, 류광렬, "병합된 예제를 이용한 자동 차 번호판 문자 인식", 한국정보과학회 2004 가을 학술발표논문집(I), 제31권, 제2호, pp. 238-240, 2004.
[22] 이경순, 안동언 "문서분류에서 가상문서기법을 이용한 성능 향상", 정보처리학회논문지, 제11-B권, 제4호, pp. 501-508, 2004.
[23] Newman, D. J., Hettich, S., Blake, C. L. and Merz, C. J., UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html], CA: University of California, Department of Information and Computer Science, Irvine, 1998.
[24] Weka3 - Data Mining with Open Source Machine Learning Software in Java http://www.cs.waikato.ac.nz/~ml/weka.
[25] Witten, I. H. and Frank, E., Data Mining-Practical Machine Learning Tools and Techniques with Java Imp.lementations, Morgan Kaufman Publishers, 1999.

이 유 정
현재 부산대학교 지능형시스템연구실의 석사과정에 재학 중이다. 부산대학교 컴퓨터공학과 학사학위(2005년)를 취득하였다. 주요 연구 관심분야는 학습, 최적화 등이다.

강 병 호
현재 부산대학교 지능형시스템연구실의 박사후 연구원으로 재직 중이다. 부산대학교 컴퓨터공학과 학사학위(1994년)를 취득한 후, 현대중공업에서 1994년에서 1997년까지 근무하였고, 부산대학교에서 컴퓨터공학과 석사학위(1999년)와 박사학위(2006년)를 취득하였다. 주요 연구분야는 최적화, 기계학습, 지능형물류시스템 등이다.

강 재 호
정보과학회논문지 : 소프트웨어 및 응용
제 33 권 제 11 호 참조

류 광 렬
정보과학회논문지 : 소프트웨어 및 응용
제 33 권 제 11 호 참조