가상예제를 이용한 수치 및 범주 속성 데이터의 분류 성능 향상

이유정[O] 강재호 강병호 류광렬
부산대학교 컴퓨터 공학과
{yjlee[O], jhkang, bhokang, krryu}@pusan.ac.kr

Improving Classification Accuracy for Numerical and Nominal Data using Virtual Examples*

Yujung Lee[O] Jaeho Kang Byoungho Kang Kwang Ryel Ryu
Department of Computer Engineering, Pusan National University

요 약

본 논문에서는 베이지안 네트워크를 기반으로 생성하고 평가한 가상예제를 활용하여 범주속성 및 수치속성 데이터에 대한 분류 성능을 향상시키는 방안을 제안한다. 가상예제를 활용하는 종래의 연구들은 주로 수치 속성 데이터를 대상으로 한 반면 본 연구에서는 범주속성 데이터에 대해서도 가상예제를 적용하여 효과를 확인하였다. 그리고 대상 도메인에 특화된 지식을 활용하여 특정 학습 알고리즘의 성능을 향상시키는 것을 목표로 한 기존 연구들과는 달리 본 연구에서는 도메인에 특화된 지식을 활용하는 대신 주어진 훈련 집합을 기반으로 만든 베이지안 네트워크로부터 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 데이터와 수치 속성을 포함한 데이터를 대상으로 한 실험 결과, 여러 가지 학습 모델의 성능이 향상됨을 확인하였다.

1. 서 론

기계 학습 분야의 연구에서는 분류 성능을 향상시키기 위한 여러 가지 노력들이 있어 왔다. 이들 중에는 특히 주어진 훈련예제 이외에 인공적으로 생성한 가상예제를 추가로 활용하여 분류 성능을 향상시키고자 하는 노력도 있었다. Marcel은 하나의 사진을 반전하거나 같은 사람의 얼굴 사진 두 장을 평균하여 가상예제를 생성하였고 이들을 얼굴 인식에 활용하여 SVM(support vector machine)의 정확도를 개선시킨 바 있다[1]. Burges와 Schölkopf는 서포트 벡터 머신을 이용한 숫자 인식 문제에서 서포트 벡터에 노이즈를 추가하여 생성한 가상 서포트 벡터를 사용하여 분류성능을 향상시켰다.[2]. Miyao는 스트로크 정보를 가진 온라인 캐릭터들에 대한 획순 등의 정보들을 활용하여 가상예제를 생성함으로써 SVM으로 오프라인 필기체 문자인식의 성능을 향상하는데 기여했다[3]. Lee와 An은 동일한 카테고리의 두 문서들을 결합하여 생성한 가상문서를 분류에 활용한 바 있다[4]. 하지만 이 연구들은 가상예제를 생성하는데 도메인 특유의 지식을 이용해야 하고, 특정한 하나의 대상 학습 알고리즘의 성능만 향상시키는 것을 목적으로 하고 있으며, 또한 수치속성 데이터에 그 사용이 한정되어 있다는 한계를 지니고 있다.

최근에는 대상 도메인에 관한 지식을 사용하지 않고 가상예제를 생성하여 사용하는 방안이 제안된 바 있다[5]. 이 방안에서는 베이지안 네트워크를 이용하여 주어진 훈련예제와 유사한 가상예제를 생성한 뒤 분류 성능 향상에 기여할 것으로 판단되는 가상예제를 선별하여 학습에 사용하였다. 이 방안에서는 또한, 분류성능을 향상시키는데 필요한 가상예제 집합의 규모를 실험적인 방법으로 통계적 유의수준을 감안하여 결정하였다. 그러나 이 방안은 범주속성의 데이터에 대해서만 적용이 가능하다는 한계를 지닌 것이다. 본 논문에서는 기존 연구의 방법을 확장하여 범주속성과 수치속성 혹은 그들이 혼합된 데이터에 대해서도 도메인 특유의 지식에 의존하지 않고 적용이 가능한 가상예제 생성 및 활용방안을 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 가상예제를 생성하고 평가하는 방안에 대하여 보다 자세히 소개한다. 이어지는 3장에서는 적절한 크기의 가상예제 집합을 결정하고 학습에 사용하는 방안을 설명한다. 그리고 4장에서는 본 제안 방안을 적용한 실험 결과를 정리하여 분석한다. 마지막 5장에서는 결론과 향후 연구로 매듭을 짓는다.

2. 유용한 가상예제 샘플링

가치 있는(useful) 가상 예제란 사전 확률(prior probability)이 높고 분류 성능 향상에 도움이 되는 가상 예제이다. 가치 있는 가상 예제를 효율적으로 얻기 위하여 이 장에서는 베이지안 네트워크를 활용하여 가상예제를 하고 평가하는 방안에 대해 자세히 설명한다.

2.1 베이지안 네트워크를 이용한 가상예제 샘플링

베이지안 네트워크는 방향성 비순환 그래프(directed acyclic graph) $G = <V, E>$와 조건부 확률 테이블들의 집합 Θ로 정의된다. 노드들의 집합 $V = \{x_1, x_2, \cdots, x_m\}$는 랜덤 변수인 속성 변수들과 클래스 x_m 변수로 구성된다. 방향 에지 $<x_j, x_i>$ E는 x_j가 x_i의 부모임을 의미한다. x_i의 부모들의 집합을 π_i라 하며, 베이지안 네트워크에서 π_i의 값이 주어지면, x_i는 나머지 변수들 $V - \pi_i - \{x_i\}$과는 조건부 독립이다. 각 노드는 조건부 확률 테이블 $\theta_i = \{p_{i,j,k} \mid j$ π_i의 가능한 값들의 집합, k x_i의 가능한 값들의 집합$\}$를 가지고 있다. $p_{i,j,k}$는 π_i의 값이 j일 때 x_i의 값이 k일 확률이다. 생성한 베이지안 네트워크가 얼마나 좋은지 평가하는 척도로 가능도(likelihood)가 널리 사용된다. 훈련 집합 $D = \{X_1, X_2, \cdots, X_n\}$에 대한 베이지안 네트워크 B의 가능도는 다음과 같이 계산된다.

여기에서, X_d는 D의 d번째 예제이며, $P_B(D)$는 B가 주어졌을 때 D의 사후 확률(posterior probability)이다. E가 결정된 상황에서 가능도가 최대화되도록 Θ를 결정하는 것은 간단하다. D에서 $\pi_i = j$, $\pi_i = j$ $x_i = k$인 예제의 수를 각각 $n_{i,j}$, $n_{i,j,k}$라 하자. sufficient statistics에 의하여 가능도를 최대화하기 위해서는 $p_{i,j,k}$를 $n_{i,j,k} / n_{i,j}$로 설정하면 된다. 최대의 가능도를 가진 베이지안 네트워크에서 샘플링을 통해 가상 예제를 생성한다면, 각 변수에 임의의 값을 설정하여 가상 예제를 생성하는 경우에 비해 사전 확률이 높은 가상 예제들을 얻을 수 있을 것이다.

베이지안 학습에서 E가 주어졌을 때 가능도가 최대화되도록 Θ를 결정하는 것은 간단하지만, 가능도를 최대화할 수 있는 E를 결정하는 것은 상대적으로 매우 어려운 문제이다. 기존 연구들에서는 고정된 구조를 사용하거나, 구조 제약하에 휴리스틱을 이용하여 구조를 결정하는 방안들이 제안되었다[6]. 하지만 이러한 방안들은 상당한 수행 시간을 필요로 한다. 따라서 본 논문에서는 속성들 간에 독립을 가정하는 구조가 고정된 나이브 베이즈를 사용하였다. 나이브 베이즈는 구조는 간단하지만, 수행 속도가 빠르고 많은 분류 문제에서 좋은 성능을 보인 바 있다.

나이브 베이즈 네트워크에서 가상예제를 샘플링하는 과정은 다음과 같다. 먼저 클래스 노드의 값을 클래스 노드안에 있는 조건 확률표를 참고하여 결정한다. 그 다음으로 각 속성값을 각자의 조건 확률표를 참고하여 결정하게 된다. 모든 속성값이 결정되면 하나의 가상예제가 생성된다.

2.2 조건부 가능도로 가상예제 평가

아무런 추가 장치 없이 베이지안 네트워크로부터 생성한 가상 예제들을 학습에 사용하면 분류 정확도가 반드시 향상될 것이라 기대할 수 없다. 학습 알고리즘이 많은 수의 가상 예제를 사용할수록 분류 성능은 가상 예제를 생성하는데 사용한 베이지안 네트워크의 분류 성능에 근접할 것이기 때문이다. 따라서 생성한 가상 예제가 분류 성능 향상에 도움이 될 지 판별할 수 있는 방안이 필요하다. 본 연구에서는 이를 위하여 "가치 있는 가상 예제는 여러 학습 알고리즘의 분류 성능 향상에 도움을 줄 것이다."라는 가설을 도입하였다. 이 가설에 따라 베이지안 네트워크의 정확도를 향상시키는 가상 예제는 다른 학습 알고리즘의 성능 향상에도 도움을 줄 것이라 예상할 수 있다. 조건부 가능도가 증가하도록 베이지안 네트워크를 변경하면 분류 정확도가 향상된다는 증명 및 실험에 관한 기존 연구들이 있다[7,8,9,10]. 이러한 기존 연구들과 위의 가설에 기반하여 베이지안 네트워크의 조건부 가능도를 증가시키는 가상 예제를 가치 있는 가상 예제라 판별한다. 조건부 가능도는 다음과 같이 정의된다.

$$CL(D|B) = \prod_d \prod_i P_B(x_{d,m} | x_{d,1}, x_{d,2}, \cdots, x_{d,m-1}) \quad (2)$$

여기서, $x_{d,m}$는 훈련 예제 X_d의 클래스 값을 의미한다. 본 제안 방안은 다음과 같이 동작한다. 먼저 (1) 주어진 훈련 집합 T로 베이지안 네트워크 B를 구성한다. (2) 가상 예제 v를 B로부터 샘플링하여 생성한다. (3) 가상 예제 v를 훈련 예제로 사용하여 B를 업데이트한다. (4) B의 T에 대한 조건부 가능도가 v를 추가하기 전보다 증가되었다면 v를 가치 있는 가상 예제로 판별하여 가상 예제 집합에 추가한다. 그렇지 않다면 B를 이전 상태로 되돌린다. (2) ~ (4)의 과정을 반복하여 필요한 수만큼 가상 예제를 얻는다.

한 개의 유용한 가상예제를 훈련집합에 추가한 후, 베이지안 네트워크의 각 조건 확률표안에 있는 엔트리들은 인클리멘탈하게 업데이트될 수 있다. 추가된 가상예제를 $v = (x_{n+1,1}, x_{n+1,2}, \cdots, x_{n+1,m})$ 라고 하자. 나이브 베이즈 네트워크에서는 모든 속성의 부모가 클래스이기 때문에 조건 확률표의 각 엔트리들은 그림 1과 같이 업데이트 된다.

$$
\begin{aligned}
&a = index_i(x_{n+1,i}) \\
&c = index_m(x_{n+1,m}) \\
&n'_{i,j,k} \leftarrow n_{i,j,k} + 1, && 1 \le i \le m,\ j = c,\ k = a \\
&n'_{i,j,k} \leftarrow n_{i,j,k}, && 1 \le i \le m,\ j = c,\ k \ne a \\
&n'_{i,j} \leftarrow n_{i,j} + 1, && 1 \le i \le m,\ j = c \\
&n'_{i,j} \leftarrow n_{i,j}, && 1 \le i \le m,\ j \ne c \\
&p_{i,j,k} \leftarrow \frac{n'_{i,j,k}}{n'_{i,j}}, && 1 \le i \le m,\ j = c,\ k = a
\end{aligned}
$$

그림 1. 가상예제 생성후 나이브 베이즈 업데이트

$index_i(x_i)$ 함수는 속성값 x_i 가 도메인 상에서 가지는 인덱스를 리턴한다. 이와 같은 인클리멘탈한 업데이트는 빠른 가상예제의 생성에 효과적이다.

2.3 수치속성이 포함되어 있는 가상예제의 생성

수치속성이 포함되어 있는 데이터에 대해서는 감독 엔트로피 이산화(supervised entropy discretization)를 먼저 수행한 후 가상예제를 생성한다. 수치속성을 이산화하게 되

면 수치속성에서 범주속성으로 변환된다. 변환한 데이터에서 2.1절과 2.2절에 소개된 과정을 통해 범주속성으로 이루어진 가상예제 생성하고 평가한 다음 유용하다고 판단된 가상예제들을 일정한 수만큼 모아 한번에 수치화하는 과정을 거친다.

수치화하는 과정은 다음과 같다. 먼저 가공하지 않은 훈련집합에서 각 속성의 전체구간을 일정한 간격으로 적당하게 잘라 그 간격에 따른 예제의 분포를 조사한다. 가상예제의 각 속성 값으로 결정된 이산화구간 내에서 하나의 일정 구간을 예제분포대로 결정한다. 결정된 일정 구간 내에서는 랜덤으로 수치를 결정하여 가상예제의 속성값으로 취한다.

3. 학습을 위한 가상예제 집합 선택

위와 같은 방법에는 두 가지 단점이 있다. 즉, 가상 예제를 몇 개나 생성하는 것이 적절한지 결정하기 어렵다는 점과 생성한 가상 예제 집합이 나이브 베이즈 이외의 학습 방법에도 도움이 된다는 것을 어떻게 보장하느냐 하는 점이다. 특히 나이브 베이즈가 다른 학습 알고리즘들에 비해 성능이 떨어지는 문제에서 이러한 방법으로 생성한 가상 예제들이 다른 학습 알고리즘의 성능 개선에 도움이 될 것이라고 예측하기란 어렵다.

이러한 문제점들을 해결하기 위하여 본 연구에서는 생성한 가상 예제들의 순서 $V = \langle v_1, v_2, \cdots, v_n \rangle$에서 여러 가지 크기의 가상 예제 집합들 $V^b = \{V_1, V_2, \cdots, V_b\}$, $V_i = \{v_j \mid 1 \le j \le i \times n / b\}$을 생성한다. 그리고 각 가상 예제 집합을 학습에 사용할 때 예상되는 정확도를 교차 검증[11]을 통해 추정하여, 가상 예제를 사용하지 않을 때에 비해 현저하게 정확도가 향상되는지 확인하다. 정확도 추정을 위한 교차 검증 시 가상 예제들은 학습에만 사용된다. 통계적 검증을 위해서는 t-검증을 사용한다[12]. t-검증 결과 신뢰도가 c 이상으로 정확도가 향상되었다고 말할 수 있는 가상 예제 집합들 중에서 평균 정확도가 가장 높은 가상 예제 집합을 최종적으로 선정한다. 어느 가상 예제 집합도 정확도가 향상된다고 확신할 수 없을 경우에는 가상 예제를 사용하지 않고 분류기를 생성한다. 본 제안 방안의 전체 알고리즘을 그림 2에 나타내었다.

4. 실험 및 결과분석

본 제안 방안을 검증하기 위하여 UCI 데이터[13] 중에서 문자 속성의 분류 문제 11가지와 수치속성이 포함되어 있는 분류 문제 9가지를 대상으로 실험하였다. 표 1은 실험에 사용한 문자속성 데이터들의 특성을 정리하였고 표 2는 수치속성을 포함한 데이터의 특성을 정리하였다.

학습에는 naive Bayes (*NB*), nearest neighbor (1-*NN*) [14], decision tree (*C*4.5) [15], NBtree[16] 다섯 가지 학습 알고리즘 방법을 적용하였다. Weka 데이터 마이닝 소프트웨어[17]를 사용하여 실험하였다. 그리고 가상 예제는 주어진 데이터의 최대 500%까지 생성하고, 50% 단위로 10개의 가상 예제 집합을 구성하였다. 각각의 실험에서 10분할 교차 검증을 수행하여 결과를 평균하였다.

표 3에 문자속성 데이터에 대한 실험 결과를 정리하였다. 표에서 *C*4.5는 가상 예제없이 학습한 경우이며, $C4.5^{+V}$는 본 제안 방안을 적용한 경우이다. 다른 학습 방법들도 동일한 방식으로 표기하였다. 본 제안 방안을 적용하여 통계적으로 의미 있는 성능 변화가 발생한 경우에는 수치 왼쪽에 화살표를 표기하였다. 이를 위하여 95% 신뢰도를 설정하여 t-검증을 수행하였다. 표에서 알 수 있듯이 *NB*의 경우 11개의 데이터 중에서 7개의 데이터에 대하여 성능 향상이 확인되었으며, $1\text{-}NN^{+V}$와 $C4.5^{+V}$는 각각 6개의 데이터에 대하여 정확도가 증가하였다. 마지막으로 *NBtree*에서는 5개의 데이터에서 정확도가 향상되었다.

표 3을 자세히 살펴보면 흥미로운 실험 결과를 확인할 수 있다. audiology의 경우 *NB*가 다른 학습 방안들에 비해 성능이 떨어지는 대표적인 문제이다. 하지만 이 문제에서 나이브 베이즈로부터 생성한 가상 예제들이 *NB*뿐 아니라 1-*NN*, *C*4.5과 *NBtree*의 분류 성능 향상에도 상당한 도움이 되었음을 알 수 있다. 이러한 결과는 3장에서 소개한 가설이 상당한 신빙성이 있음을 의미한다.

표 4은 수치속성을 포함한 데이터에 대한 실험 결과를 정리하였다. 문자속성으로만 이루어진 데이터와 마찬가지로 평균적으로 정확도 향상에 본 논문에서 생성한 가상예제들이 훈련예제의 분류성능 향상에 좋은 영향을 끼친 것을 알 수 있다.

나이브 베이즈와 표면적으로 관계가 없어 보이는 1-*NN*, *C*4.5에서도 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.

5. 결론

본 논문에서는 베이지안 네트워크를 통해 생성한 가상 예제를 활용하여 분류 성능을 높이는 방안을 소개하였다. 본 제안 방안은 베이지안 네트워크로부터 훈련집합과 유사한 가상 예제를 생성하고 그 적합도를 조건부 가능도로 평가하였다. 통계적 검증 실험을 통하여 적용하고자 하는 학습 알고리즘에 적절한 가상 예제 집합의 규모를 결정하여 주어진 훈련 집합과 함께 학습에 사용하였다. 범주속성 및 수치속성 데이터들을 대상으로 실험한 결과 본 제안 방안이 여러 학습 알고리즘의 성능을 개선하는데 도움이 됨을 확인하였다.

앞으로의 연구 방향은 본 연구에서 사용한 가장 간단한 베이지안 네트워크 구조인 나이브 베이즈 대신 보다 높은 분류 정확도를 보인다고 알려진 TAN과 같은 베이지안 네트워크를 사용한다면 보다 좋은 품질의 가상 예

제를 생성할 수 있을 것이라고 예상된다. 그리고 생성된 가상예제들이 훈련예제에 어떤 영향을 주는지에 대한 분석적 후속 연구가 필요하다.

표 1. 문자속성으로 이루어진 데이터의 특성

Data set	# of examples	# of attributes	# of classes
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

표 2. 수치속성을 포함한 데이터의 특성

Data set	# of examples	# of attributes	# of classes	# of numberic attributes
lymph	148	18	4	3
iris	150	4	3	4
hepatitis	155	19	2	6
sonar	208	60	2	60
glass	214	9	7	9
heart-c	303	13	5	6
labor	57	16	2	8
heart-statlog	270	13	2	13
ionosphere	351	34	2	34

표 3. 문자속성으로 이루어진 데이터에 대한 가상예제 추가 실험결과

Data set	*NB*	NB^{+V}	1-*NN*	1-NN^{+V}	*C*4.5	*C*4.5^{+V}	*NBtree*	*NBtree*$^{+V}$
audiology	73.0	⬆ 79.2	79.0	⬆ 81.8	77.5	⬆ 83.2	78.3	⬆ 82.3
breast-c	73.0	73.6	72.0	70.5	75.2	74.2	71.8	⬆ 73.8
kr-vs-kp	87.7	⬆ 89.7	90.0	90.8	99.4	99.2	97.1	97.8
monks-1	73.4	72.2	79.9	79.0	80.7	80.1	90.6	89.4
monks-2	59.7	59.3	59.2	⬇ 57.2	62.2	⬇ 60.6	60.4	⬇ 59.2
monks-3	93.3	93.3	84.6	⬆ 93.3	91.1	⬆ 92.2	93.4	93.4
primary	48.4	⬆ 50.8	38.9	⬆ 40.0	41.3	⬆ 43.5	46.0	⬆ 47.6
soybean	92.2	⬆ 93.5	91.6	⬆ 92.9	92.1	⬆ 93.8	91.5	⬆ 92.5
splice	95.3	⬆ 96.1	75.9	⬆ 82.4	94.1	⬆ 95.8	95.3	95.0
vote	90.1	⬆ 92.8	92.6	⬆ 93.2	96.8	97.1	95.6	⬆ 96.8
zoo	93.0	⬆ 93.8	96.0	96.0	93.1	⬆ 95.0	95.1	⬇ 94.4

표 4. 수치속성을 포함한 데이터에 대한 가상예제 추가 실험결과

Data set	*NB*	NB^{+V}	1-*NN*	1-NN^{+V}	*C*4.5	*C*4.5^{+V}	*NBtree*	*NBtree*$^{+V}$
lymph	83.8	⬆ 85.1	79.0	79.7	78.3	77.0	78.2	⬆ 79.7
iris	95.3	96.0	95.3	⬆ 96.7	94.0	⬆ 95.3	94.0	94.0
hepatitis	83.7	⬆ 85.2	82.0	81.4	78.7	78.1	82.6	⬇ 80.1
sonar	67.8	⬆ 71.7	85.5	85.5	72.6	⬆ 76.5	77.8	78.4
glass	45.8	⬆ 46.8	70.5	⬇ 68.7	69.2	70.5	68.7	⬆ 69.7
heart-c	83.2	83.6	75.3	75.3	76.6	76.6	78.9	⬆ 81.3
labor	86.0	⬆ 88.2	82.5	83.2	73.7	⬆ 76.3	87.7	⬆ 89.3
heart-statlog	81.1	⬆ 83.5	75.2	⬆ 77.5	76.7	79.3	78.9	79.8
ionosphere	89.2	⬆ 91.5	86.3	⬆ 88.1	91.5	⬆ 90.1	89.7	90.5

```
procedure Learning-with-Virtual-Examples
  input
    L – a learning algorithm to use
    T – a training set
    n – a maximum number of virtual examples to be generated
    b – a number of virtual sets derived
    c – a confidence of t-test for selecting a virtual set

  output
    h – a classifier derived by L with virtual examples

  begin
    V ← ø
    Build a Bayesian network B using T
    Calculate the conditional likelihood cl of T on B
    repeat until |V| < n
      Generate a virtual example v by sampling from B
      Copy B to B+ and incrementally update B+ with v
      Calculate the conditional likelihood cl+ of T on B+
      if cl+ > cl then
        Add v to V
        B ← B+, cl ← cl+
      end if
    end repeat

    Generate V* = {V1, V2, ..., Vb}, Vi  {vj | 1 ≤ j ≤ i × ⌈n / b⌉}.
    Get accuracy list A0 by using L with T (ten cross-validations are used)
    a* ← average of A0, V* ← ø
    foreach Vi in V*
      Get accuracy list Ai by using L with T ∪ Vi
      if Ai is higher than A0 by t-test with confidence c then
        if average of Ai is greater than a* then
          a*  average of Ai, V*← Vi
        end if
      end if
    end foreach

    Derive a classifier h by using L with T ∪ V*
    return h
  end
end procedure
```

그림 2. 가상예제 생성 알고리즘

감사의 글

이 논문은 한국과학재단 국가지정연구실사업의 지원으로 이루어진 것임 (Contact number : M10400000279 - 05J0000 - 27910)

참고문헌

1. Marcel, S, "Improving Face Verification using Symmetric Transformation," Proceedings of IEICE Transactions on Information and Systems, Vol. E81, No. 1, pp. 124-135,2003.
2. Burges, C. and Schölkopf, B., "Improving the Accuracy and Speed of Support Vector Machines," Advances in Neural Information System, Vol. 9, pp. 375-381, 1997.
3. Miyao, H., Maruyama, M., Nakano, Y., Hanamoi, T.: Off-line handwritten character recognition by SVM based on the virtual examples synthesized from on-line characters. Proc. of Eighth International Conference on Document Analysis and Recognition. Vol. 1 (2005) 494-498
4. 이경순, 안동언 "문서분류에서 가상문서기법을 이용한 성능 향상," 정보처리학회논문지, 제11-B권, 제4호, pp. 501-508, 2004.
5. Y. Lee, J. Kang, B. Kang, K. R. Ryu "Sampling of Virtual Examples to Improve Classification Accuracy for Nominal Attribute Data", The Fifth International Conference on Rough Sets and Current Trends in Computing (RSCTC), 2006, 게재예정
6. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. Machine Learning, 29:131--163, 1997.
7. 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.
8. 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.
9. Burge, J. and Lane, T., "Learning Class-Discriminative Dynamic Bayesian Networks," Proc. of the 22th International Conference on Machine Learning, pp. 97-104, 2005.
10. Pernkopf, F. and Bilmes, J., "Discriminative versus Generative Parameter and Structure Learning of Bayesian Network Classifiers," Proc. of the 22nd International Conference on Machine learning, pp. 657-664, 2005.
11. Bouckaert, R. and Frank, E., "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms," Proc. of the 8th Pacific-Asia Conference, pp. 3-12, 2004.
12. Wild, C.J, and G. A. F. Seber, Introduction to probability and statistic, Department of Statistics, University of Auckland, 1995.
13. 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.
14. Aha, D. and Kibler, D., "Instance-based Learning Algorithms," Machine Learning, Vol.6, pp. 37-66, 1991.
15. Quinlan, J. R., C4.5 : Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.
16. Ron Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: a decision tree hybrid", Proc. of the Second International Conference on Knowledge Discovery and Data Mining, 1996.
17. Witten, I. H. and Frank, E., Data Mining-Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufman Publishers, 1999.