Proceedings of KIIS Spring Conference 2008 Vol. 18, No. 1.

정보이론과 신경망의 가중치를 이용한 속성선택

Feature Selection Algorithm using Information theory and Neural Networks

조재훈[1], 이대종[2], 전명근[1]

[1]충북 청주시, 충북대학교 전기전자컴퓨터공학부
[2]충북 청주시, 충북대학교 BK21 충북정보기술사업단
E-mail : mgchun@chungbuk.ac.kr

요 약

본 논문에서는 신경망의 가중치와 정보이론을 이용한 속성선택 기법을 제안하였다. 제안된 방법은 정보이론의 상호정보량을 이용하여 각 속성들의 중요도를 평가한 후 중요도가 높은 속성들만을 선택하여 신경망의 입력으로 사용한다. 신경망의 입력으로 선택된 속성의 가중치에 대한 평가를 통하여 오차에 큰 영향을 미치는 속성들을 순차적으로 제거하여 가장 우수한 속성들을 구한다. 제안된 기법의 성능을 평가하기 위하여 다양한 패턴 분류 문제에 적용하고 그 성능이 우수함을 확인하였다.

Key Words : 정보이론, 신경망의 가중치 속성선택 기법

1. 서 론

속성선택기법은 원 데이터에서 중요도가 높은 속성만을 선택하여 데이터의 신뢰성을 높이는 처리과정으로서 데이터 마이닝, 패턴 인식, 기계 학습 등의 다양한 분야에서 중요한 분야로 인식되어져 왔다. [1].

속성선택 방법은 평가 조건들에 따라서 크게 필터(Filter) 방법, 랩퍼(Wrapper) 방법 그리고 혼합형(Hybrid) 방법으로 나눠질 수 있다. 랩퍼 기법에 기반을 둔 속성선택기법은 반복적인 속성들의 선택을 통한 분류기의 성능 평가로서 속성들을 선택하기 때문에 일반적으로 필터(filter) 기법으로 불리는 속성선택기법보다 우수한 성능을 보이는 장점이 있는 반면에 필터 기법에 비하여 연산속도가 느린 단점을 가지고 있다. 혼합형 방법은 랩퍼 기법과 필터 기법을 융합하거나 다른 기법들과 융합한 기법으로 각각의 기법들의 단점을 보완하여 최적의 속성부분집합을 선택하는 방법으로서 현재 다양하게 연구되어지고 있다[2].

본 논문에서는 정보이론의 상호정보량과 신경망의 가중치를 이용하여 각각의 속성들을 평가하여 최적의 속성부분집합을 선택하는 방법을 제안하였다. 제안된 기법의 유용성과 성능을 평가하기 위하여 UCI Machine-Learning Repository [3]데이터에 적용하고 기존 기법들의 결과들과 비교하여 그 타당성을 보이고자 한다.

2. 정보이론을 이용한 속성선택 기법

속성선택에서 속성 자체와 속성이 속하는 클래스를 각각의 확률변수로서 처리할 수 있다. 정보이론에서 데이터의 각 속성은 확률변수로 볼수 있으며, 엔트로피로 정의될 수 있다. 랜덤변수 X 가 $p(x)=\Pr\{X=x\}, x\in\lambda$의 소스알파벳 λ을 가진다면 X의 엔트로피는 아래 식(1)과 같이 계산된다.

$$H(X)=-\sum_{x\in\lambda}p(x)\log p(x) \qquad (1)$$

본 논문에서는 필터 기반의 전처리 선택으로서 엔트로피를 이용한 상호정보량의 특징을 이용하여 각각의 속성들과 클래스 간의 상호정보량을 계산하고 각 속성들의 중요도를 평가하였다.

3. 신경망 가중치를 이용한 속성선택

신경회로망은 문제의 입력들이 올바른 출력을 생성하도록 연결 강도의 값을 조정한다. 즉, 출력층에서 계산된 오차에 의해서 각 입력값에 연결된 가중치를 조정하여 오차가 최소가 되도

록 학습시킨다. 그러나 각 가중치 변화는 활성함수의 종류에 따라 복잡하고 분석이 어렵다는 단점들을 가진다. 이러한 문제점들을 해결하기 위하여 옥닉 층과 출력 층의 근사적인 이득을 이용하는 방법이 연구되어졌다[4].

$$LG_{ik} = \sum_{j} | W_{ij} \times W_{jk} | \qquad (2)$$

$$ANNIGMA_{ik} = \frac{LG_{ik}}{\max(LG_k)} \times 100 \qquad (3)$$

4. 신경망 가중치와 상호정보량을 이용한 속성선택

제안된 기법은 크게 필터기법단계와 랩퍼기법 단계로 나눠지고 단계별로는 아래와 같이 수행되어진다.

[필터기법 단계]

[단계 1] 각 속성들에 대한 상호정보량을 계산한다.

[단계 2] 데이터의 크기에 따라 입력데이터의 모든 속성을 선택 또는 상호정보량의 큰 순위에 따라 우수한 속성만을 선택한다.

[랩퍼기법 단계]

[단계 1] 필터 단계에서 선택된 속성 집합에서 속성부분집합을 생성한다.

[단계 2] 학습데이터에서 검증데이터를 이용하여 학습된 신경망의 성능을 평가한다.

[단계 3] 단계 2에서 계산된 신경망의 오차와 단계 1에서 학습된 가중치를 기반으로 각각의 속성들에 대한 MIIGMA를 계산한다.

[단계 4] MIIGMA의 순위를 기반으로 속성부분집합을 선택하고 종료 조건을 판별하여 종료하거나 반복 연산을 위해 수행한다.

단계 3에서 정보이론의 상호정보량에서 평가된 속성들의 유효성을 랩퍼기법의 가중치 이득에 반영하기 위하여 아래의 수식으로 정의하고 평가하였다.

$$MIIGMA_{ik} = \left(\frac{LG_{ik}}{\max(LG_k)} + \frac{MI_i}{\max(MI_i)} \right) \times 100 \qquad (4)$$

위 수식 (4)에서 MI는 각각의 속성들에 대한 상호정보량을 의미한다.

4. 시뮬레이션 및 결과 고찰

제안한 기법의 성능을 평가하기 위하여 UCI 데이터들을 사용하였다. 실험에 사용된 데이터들은 학습데이터들과 검증데이터로 구분되어져 있다. 제안된 기법을 이용하여 속성들을 평가할 때는 원 학습데이터를 10-fold로 나누고 9개 fold는 학습데이터로 1개의 fold는 검증데이터로 재구성하여 속성선택에 사용하였으며, 최종적으로 선택된 속성 집합들과 원 검증데이터를 이용하여 성능을 평가하였다. 표 1에서 제안된 방법과 다른 속성부분집합 선택 기법들과의 성능을 비교하였다. 비교 결과 값들은 참고문헌 [4]의 비교 결과 값을 사용하였다. 표 1에서 알 수 있듯이 제안된 기법이 다른 기법들에 비해 대분분 우수한 성능을 보임을 알 수 있었다 .

5. 결 론

중복성이 적고 유용한 속성들은 분류기나 예측기의 성능을 신뢰성 있게 만들 수 있고, 반면에 중복성, 노이즈 및 유용성이 적은 속성들은 출력에 큰 오차를 유발시킬 수 있다. 이러한 오차를 줄이고 신뢰성들을 확보하기 위해서는 속성들의 특성이나 유용성들을 평가하여 오차에 기여하는 속성이나 잘못된 데이터들을 제거할 필요가 있다. 본 논문에서는 신경망의 가중치들와 정보이론을 융합한 속성선택기법을 제안하였다. 제안된 방법을 UCI 데이터를 이용하여 평가하였으며, 기존의 다른 기법들보다 우수한 결과를 보임을 확인할 수 있었다.

1. 제안된 기법의 성능비교

데이터	ANNIGMA-Wrapper		MIIGMA-Wrapper	
	속성 수	오차(%)	속성 수	오차(%)
3P	3.0±0.0	0.0±0.0	3.0±0.3	0.0±0.0
Monk3a	2.3±0.7	2.9±0.8	2.2±0.4	3.1±0.2
Monk3b	2.2±0.4	2.8±0.0	2.4±0.3	2.4±0.3
Cancer	5.8±1.3	3.5±1.2	6.1±1.3	2.5±0.1
Credit	6.7±2.5	12.0±0.8	6.0±2.7	9.9±1.2
Heart(LB)	2.7±1.2	22.3±2.0	2.6±0.9	17.4±1.5
Ionosphere	9.0±2.5	9.8±1.3	9.2±1.3	9.3±1.4
Pima	5.2±1.4	22.2±1.4	4.9±1.7	11.2±0.7
Vote	3.3±1.9	3.1±0.2	2.7±1.6	2.5±0.8

참 고 문 헌

[1] M.Dash. and H.Liu, "Feature selection for classification," Intelligent Data· Analysis, Vol 85, Isuue 2, pp 234-247, 1997.

[2] M. Dash and H. Liu, "Hybrid search of feature subsets," Proc. PRICAI, singapore, 1998

[3] C. J. Merz and P. M. Murphy. UCI repository of machine learning databases. Dept. Computer Science, Univ. California, Irvine. [Online].

[4] C.N. Hsu, H.J. Huang, and D. Schuschel, "The ANNIGMA-Wrapper Approach to Fast' Feature Selection for Neural Nets,"IEEE Trans. on Syst. man and Cybernetics-PART B: CYBER., V. 32, NO. 2, 2002