Proceedings of KFIS Autumn Conference 2007 Vol. 17, No. 2.

유전자 알고리즘과 정보이론을 이용한 속성선택

Feature Selection by Genetic Algorithm and Information Theory

조재훈[1], 이대종[2], 송창규[2], 전명근[1]

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

요 약

속성선택(Feature Selection)은 패턴분류 문제에서 분류기들의 성능을 향상시킬 수 있는 중요한 부분으로 다양한 기법들이 연구되어지고 있다. 특히, 많은 변수와 속성들을 가지는 데이터를 패턴분류 하는 과정에서 주요 속성부분집합을 추출하여 이용함으로써 분류기의 연산속도 및 정확도를 향상시킬 수 있다. 본 논문에서는 유전자 알고리즘과 정보이론의 상호정보량을 이용하여 속성선택을 하는 기법을 제안하였다. 제안된 기법의 성능을 평가하기 위하여 패턴분류 문제에 적용하고 그 성능이 우수함을 확인하였다.

Key Words : 속성선택, 패턴분류, 유전자 알고리즘, 상호정보량

1. 서 론

속성선택(Feature Selection)은 주어진 문제에서 효과적이고 개선된 해를 얻기 위해 유용한 속성들을 선택하는 처리과정이다. 일반적으로 모든 이용 가능한 속성들이 분류기의 좋은 성능에 기여하지는 않는다. 속성들 중에는 서로 다른 속성들 사이에서 혼란을 야기 시키는 중복성이나 분류기 성능을 저하 시키는 잡음들을 가지고 있는 속성들이 존재하기 때문에 주어진 문제를 해결하기 위해 사용될 수 있는 데이터 집합으로부터 최적의 속성부분집합을 선택하는 것은 패턴 분류 문제에서 중요하게 인식되어지고 있다.

속성선택기법은 검색 특징을 기반으로 exhaustive 기법, heuristic 기법 그리고 random 기법으로 나눠질 수 있다[1]. 또한, 속성들의 선택에 이용되는 성능 평가 함수들을 기반으로 distance [2], information [3], dependence [4], consistency [5], 그리고 분류기의 오차비율을 이용하는 5개의 기법으로 구분할 수 있다.

속성선택의 조건으로서 분류기 오차 비율을 사용한 속성선택은 랩퍼(wrapper) 기법으로 알려져 있다. 랩퍼 기법에 기반을 둔 속성선택기법은 반복적인 속성들의 선택을 통한 분류기의 성능 평가로서 속성들을 선택하기 때문에 일반적으로 필터(filter) 기법으로 불리는 속성선택 기법보다 우수한 성능을 보이는 장점이 있다. 반면에, 랩퍼 기법은 필터 기법에 비하여 연산속도가 느린 단점을 가지고 있다. 필터 기법에 기반을 둔 속성 선택은 학습알고리즘에 독립적으로 행해지고, 전처리 과정에서 적절한 속성들을 선택하기 때문에 연산시간이 빠르다.

속성선택기술은 다양한 분류기에 적용되어져 왔다. Mao는 가지치기(pruning) 분석과 SVM (support vector machine)을 이용한 속성선택을 제안하였고[6], Hsu 등은 신경회로망에서 가중치를 기반으로 하는 ANNIGMA(artificial neural net input gain measurement approximation)-랩퍼 기법을 제안하였다[7]. Pal 과 Chntalpudi는 신경회로망을 학습하는 동안 중요 속성을 온라인으로 선택하는 개선된 기법을 제안하였다[8]. 또한, 무작위 기법으로서 진화 알고리즘의 속성선택 적용에 대한 연구들도 시도되어져왔다. 일반적으로, 유전자 알고리즘 기반 속성선택 기법에서 집단의 각각의 개체(염색체)들은 속성의 부분집합으로 표현한다.

n-차원 속성 공간에 대해서 각각의 염색체는 n-bit 이진스트링으로 표현되고, i번째 속성이 염색체에 의해 1로 표현되면, 그 속성은 속성부분집합으로서 선택되어지는 기법들이 일반적으로 사용되어왔다. [9]에서, 유전자 알고리즘 기반 속성선택기술이 고차원데이터에 대해

다양한 고전적인 속성선택 기술보다 더 우수한 성능을 가지는 것을 보였으며, Foroutan 와 Sklansky 는 유전자 알고리즘을 이용한 속성선택에 대해 분기한정법(branch and bound) 기술을 이용하였다[10]. Pal등은 속성선택에 대해 self-crossover로 불리는 새로운 유전 연산자를 제안하였다[11]. 현재까지도 다양한 진화 알고리즘을 이용한 속성선택 기법들이 연구되어지고 있다.

본 논문에서는 랩퍼와 필터 기법의 두 특징을 가지는 융합된 구조의 속성선택기법을 제안하였다. 먼저 상호정보량과 필터 기법을 이용하여 속성들을 전처리하여 우수한 속성들을 선택한 후 유전자 알고리즘과 신경회로망을 이용하여 랩퍼 기법으로 가장 적절한 속성들을 선택한다. 제안된 기법의 유용성과 성능을 평가하기 위하여 UCI Machine-Learning Repository [12]데이터에 적용하고 기존 기법들의 결과들과 비교하여 그 타당성을 보이고자 한다.

2. 상호정보량을 이용한 속성선택

속성선택문제에서 유효한 속성들은 출력에 대하여 중요한 정보들을 많이 포함하고, 반대로, 그렇지 못한 속성들은 작은 량의 정보들만을 포함한다. 분류 문제를 해결하기 위해서는 입력 속성에서 가능한 한 많은 정보들을 포함하도록 속성들을 선택 하여야한다. 분류 문제에서 속성을 F, 클래스를 C라고 하면 속성과 클래스간의 상호정보량은 다음과 같이 정의 된다.

$$I(F;C)=\iint P(f,c)\log\frac{P(f,c)}{P(f)P(c)}dxdy \qquad (1)$$

상호정보량이 클수록 속성이 클래스에 대해 많은 정보를 포함하게 된다. 이런 특성을 이용하여 각각의 속성들을 해당 클래스에 대해 상호정보량을 계산하고 그 순위들을 보면 속성들의 중요도를 어느 정도 판단할 수 있게 된다.

본 논문에서는 이런 특성을 이용하여 래퍼 기반 속성선택 전의 전처리 과정으로서 속성들의 상호정보량을 계산하여 순위를 기반으로 속성들의 수를 줄였다.

3. 유전알고리즘을 이용한 속성선택

유전자 알고리즘의 기본 구조는 크게 초기화, 적합도 평가와 재생산, 교배, 돌연변이의 4단계로 구분된다. 초기화 단계에서는 최적화 문제의 해가 될 가능성이 있는 개체들의 집단이 형성된다. 해가 될 수 있는 해공간상의 초기점들은 무작위로 분포 되도록 선택되거나 아니면 경험적인 기법으로 선택된다. 경험적인 초기집단 생성은 문제마다 다르기 때문에 해공간에 대한 정보가 없을 경우에는 거의 사용하지 않는다. 그 다음 단계는 적합도 평가이다. 구해진 해들은 목적함수를 제공하고 이로부터 적합도를 평가한다. 적합도가 우수한 개체들은 다른 개체들보다 더 많이 선택되어 교배를 통해 재결합되는데 서로간의 유전정보를 교환함으로써 집단에 새로운 개체를 도입하게 되고 현 집단 내에 존재하는 정보만을 이용하여 변화를 시도하게 된다.

3.1 유전자 알고리즘에서의 속성 표현

유전자 알고리즘에서의 속성들은 염색체로서 이진 스트링으로 표현된다. 각각의 비트는 하나의 속성들을 표시하고 1이면 그 속성을 선택하고 0이면 선택하지 않는다. 또한 염색체의 길이는 입력데이터의 총 속성 수와 동일한 크기로 코딩된다.

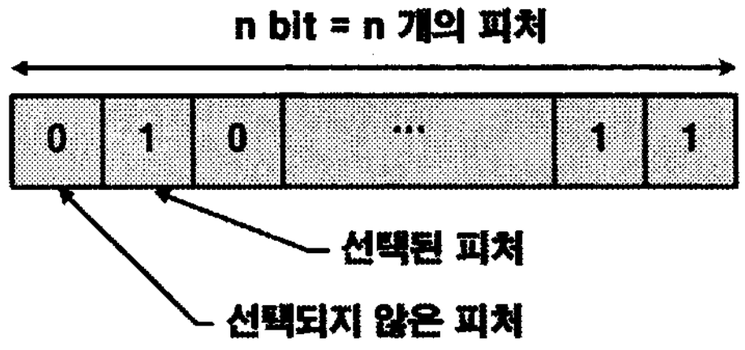

그림 1. n차원 이진 염색체

3.2 적합도 계산

속성 선택에서 유전자 알고리즘의 적합도 계산은 분류기의 성능과 선택된 속성들의 수들을 이용하는 기법이 일반적이다. 두 조건들 사이에는 설계자에 의해 비율이 조정되는 가중치들을 사용하기도 한다.

[13]에서는 위에서 설명한 적합도 평가의 조건를 아래 수식(2)로 정의하였다.

$$F=w*c(x)+(1-w)*(1/s(x)) \qquad (2)$$

위 식에서 F는 적합도, w는 분류기의 성능과 선택된 속성사이의 가중치, $c(x)$는 분류기의 성능(분류기의 정확도), $s(x)$는 선택된 속성들의 수를 나타낸다. [14]에서는 아래의 수식을 이용하였다.

$$fitness(z)=\lambda \cdot acc(z)-(1-\lambda)\cdot\frac{feats(z)}{totalfeat} \qquad (3)$$

식(3)에서 fitness(z)는 선택된 속성부분집합 z에서의 적합도, λ는 두 척도의 가중치, $acc(z)$는 분류기의 정확도, $feats(z)$는 선택된 속성 수, $totalfeat$는 입력데이터의 총 속성 수를 나타낸다. 위 두 수식에서 설계자가 두 조건을 만족하는 속성들을 선택하고자 한다면, 즉, 속성들의 수가 가장 적으면서 분류기의 성능이 가장 우수한 속성부분집합을 선택하고자 한다면 첫째항과 두 번째 항의 비율을 동일하게 적용하여야 한다.

4. 상호정보량과 유전알고리즘을 이용한 속성선택 알고리즘

그림 2에서는 본 논문에서 제안한 기법의 순서도를 나타냈다. 단계 1에서 문제의 입력데이터를 각각의 속성에 대하여 상호정보량을 계산하고, 단계 2에서는 계산된 상호정보량을 기반으로 순위를 결정한다. 순위가 높은 순서대로 다시 정리하여 상호정보량이 낮은 속성들을 제거한 후, 유전 알고리즘의 초기 집단을 생성하기 위한 후보 속성들만을 저장한다.

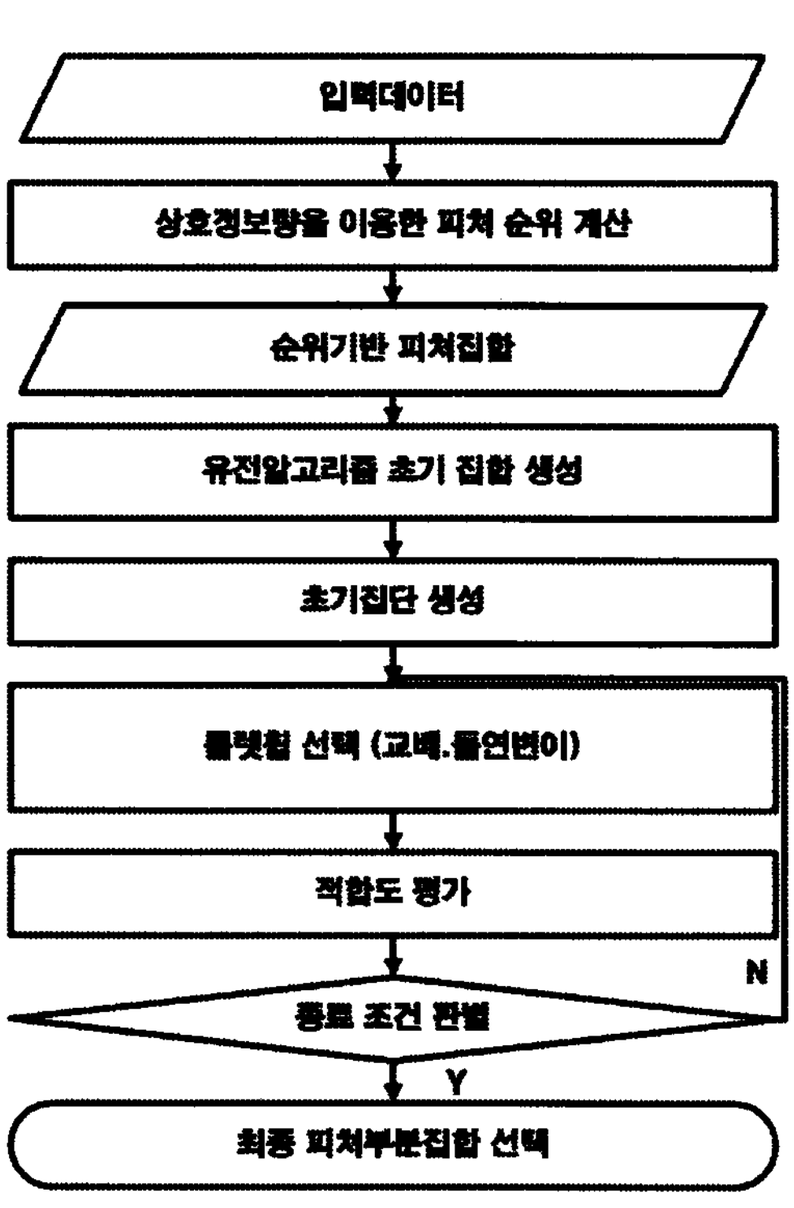

그림 2. 제안된 기법의 순서도

이 과정까지가 필터 형태의 속성선택과정이다. 단계 3에서는 단계 2에서 선택된 후보속성들을 가지고 초기 집단을 생성한다. 단계 4에서 유전 알고리즘의 선택 과정, 교배, 돌연변이를 통하여 속성들을 재생산한다. 단계 5에서는 적합도 함수를 이용하여 각각의 염색체(선택된 속성들)에 대한 평가를 수행한다. 단계 6에서는 반복횟수, 속성들의 수, 분류기의 정확도 등의 종료조건들의 판별하여 종료를 하거나 단계 4에서부터 반복적으로 속성선택과정을 수행한다.

5. 시뮬레이션 및 결과 고찰

제안한 기법의 성능을 평가하기 위하여 UCI의 기계학습 연구에서 제공하는 연구용 데이터들 사용하여 다른 기법들과 비교하였다. 실험환경은 Pentium4 1.6GHz, 1G 메모리에서 수행하였다.

표 1은 본 실험에서 사용된 유전자 알고리즘의 파라미터들을 나타냈다. 실험에서 적합도 함수는 식(3)을 이용하였으며 모든 실험에서 사용된 가중치 λ는 0.5로 설정하였다. 유전알고리즘의 선택 기법은 룰렛휠 기법을 이용하였고, 일점교배와 단순 돌연변이를 사용하였다. 분류기로는 신경회로망을 이용하였다. 표 2에서는 제안된 알고리즘과 기존의 다른 속성부분집합 선택 기법들과의 성능을 비교하였다. 선택된 속성수 항목에서는 평균 속성선택수와 표준오차로 나타냈으며, 오차 항목에서는 평균오차와 표준편차로 표시하였다. 특히, cancer 와 Ionosphere 데이터의 오차는 기존의 방법들에 비해 약 50[%]의 성능향상을 보였으며, 다른 데이터들의 오차는 평균 18[%] 정도의 성능향상을 보였다. 제안된 기법의 오차 항목에서 표준 편차가 평균 0.0으로 나타나는 이유는 유전자 알고리즘이 빠른 세대에서 최적해로 수렴되었기 때문이라 분석된다.

표 1. 유전알고리즘의 파라미터

파라미터	값
집단크기	20
교배 확률	0.7
돌연변이 확률	0.1
세대수(반복횟수)	20
가중치(λ)	0.5

6. 결 론

속성선택은 분류기의 성능을 높이기 위해 초기 입력데이터 속성들에서 유용한 속성들만을 선택하여 사용하는 기법으로 다양한 기법들이 제안 되어져 왔다. 필터 기법 기반 속성선택은 연산시간은 우수하지만 랩퍼 기법에 비하여 성능이 저하되는 단점을 가지고 있고, 랩퍼 기법 기반 속성선택은 반복적인 성능 평가로 인해

표 2. 제안된 기법에 의한 속성선택과 다른 기법들의 성능비교[7]

데이터	신경회로망		Conventional Wrapper		ANNIGMA-Wrapper		제안된 기법	
	원 속성수	오차(%)	선택된 속성수	오차(%)	선택된 속성 수	오차(%)	선택된 속성 수	오차(%)
Monk3a	6	10.0±5.2	3.4±1.6	5.1±3.4	2.3±0.7	2.9±0.8	2.0±0.3	2.2±0.0
Monk3b	15	2.8±0.0	4.4±1.1	2.8±0.0	2.2±0.4	2.8±0.0	2.2±1.0	2.2±0.0
Cancer	9	4.1±4.7	7.2±1.2	3.6±1.1	5.8±1.3	3.5±1.2	2.1±0.4	1.4±0.0
Credit	9	14.1±1.7	13.4±1.0	14.4±0.8	6.7±2.5	12.0±0.8	3.1±0.5	10.0±0.0
Ionosphere	34	11.4±3.9	32	10.2	9.0±2.5	9.8±1.3	8.2±2.3	4.2±0.0
Pima	8	24.1±5.0	6.9±1.0	23.0±1.3	5.2±1.4	22.2±1.4	3.4±0.5	18.2±0.0

성능은 우수하지만 연산시간을 증가시키는 단점을 가지고 있다.

본 논문에서는 랩퍼와 필터 기법의 두 특징을 가지는 융합된 구조의 속성선택 기법을 제안하였다. 필터 선택 단계에 상호정보량을 이용하여 각 속성들의 순위를 결정하였으며, 유전자 알고리즘과 신경회로망을 이용하여 랩퍼 기법으로 가장 적절한 속성들을 선택하였다. 제안된 기법의 성능을 평가하기 위해 UCI 데이터에 적용하고 기존의 기법들과 비교하였다. 실험결과에서 제안된 기법이 기존의 기법보다 속성선택의 수는 작으면서 분류기의 성능이 우수함을 보였다.

본 연구는 산업자원부의 지원에 의하여 기초전력연구원 (R-2007-2-046) 주관으로 수행된 과제임

참 고 문 헌

[1] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., Vol. 1, No. 3, pp. 131-156, 1997.

[2] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature selection," IEEE Trans. Comput., Vol. C-26, No. 9, pp. 917-922, Sep. 1977.

[3] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," IEEE Trans. Neural Netw., Vol. 13, No. 1, pp. 143-159, Jan. 2002.

[4] A. N. Mucciardi and E. E. Gose, "A comparison of seven techniquesfor choosing subsets of pattern recognition," IEEE Trans. Comput., Vol. C-20, pp. 1023-1031, Sep. 1971.

[5] M. Dash and H. Liu, "Consistency-based search in feature selection," Artif. Intell., Vol. 151, pp. 155-176, 2003.

[6] K. Z. Mao, "Feature subset selection for support vector machines through discriminative function pruning analysis," IEEE Trans. Syst., Man, Cybern. B, Vol. 34, No. 1, pp. 60-67, Feb. 2004.

[7] Chun-Nan Hsu, Hung-Ju Huang, and Dietrich Schuschel, "The ANNIGMA-Wrapper Approach to Fast Feature Selection for Neural Nets," IEEE Trans. on Syst. man and Cybernetics-PART B: CYBERNETICS, Vol. 32, No. 2, 2002.

[8] N. R. Pal and K. Chintalapudi, "A connectionist system for feature selection," Neural, Parallel, and Sci. Comput., Vol. 5, pp. 359-381, 1997.

[9] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," Patt. Recognit., Vol. 33, pp. 25-41, 2000.

[10] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for largescale feature selection," Patt. Recognit. Lett., Vol. 10, pp. 335-347, 1989.

[11] N. R. Pal, S. Nandi, and M. K. Kundu, "Self-crossover: A new genetic operator and its application to feature selection," Int. J. Syst. Sci., Vol. 29, No. 2, pp. 207-212, 1998.

[12] C. J. Merz and P. M. Murphy. UCI repository of machine learning databases. Dept. Computer Science, Univ. California, Irvine. Online available : http://www.ics.uci.edu/~mlearn/MLRepository.html.

[13] F. Tan, X. Fu, Y. Zhang and Anu G. Bourgeois, "Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data", IEEE Congress on Evolutionary Computation, pp. 2529-2534, 2006.

[14] J. J Aguilera, M chica, M. J. del Jesus and F. Herrera, "Niching genetic feature selection algorithms applied to the design of fuzzy rule-based classification systems", IEEE International conference on Fuzzy Systems Fuzz-IEEE2007, pp. 1-6, 2007.