한국지능시스템학회 논문지 2009, Vol. 19, No. 2, pp. 191-196

상호정보량과 Binary Particle Swarm Optimization을 이용한 속성선택 기법

Feature Selection Method by Information Theory and Particle Swarm Optimization

조재훈* · 이대종* · 송창규** · 전명근+**

Jae Hoon Cho*, Dae Jong Lee*, Chang Kyu Song** and Myung Geun Chun+**

* 충북 청주시, 충북대학교 전기전자컴퓨터공학부

** 충북 청주시, 충북대학교 BK21 충북정보기술사업단

요 약

본 논문에서는 BPSO(Binary Particle Swarm Optimization)방법과 상호정보량을 이용한 속성선택기법을 제안한다. 제안된 방법은 상호정보량을 이용한 후보속성부분집합을 선택하는 단계와 BPSO를 이용한 최적의 속성부분집합을 선택하는 단계로 구성되어 있다. 후보속성부분집합 선택 단계에서는 독립적으로 속성들의 상호정보량을 평가하여 순위별로 설정된 수만큼 후보속성들을 선택한다. 최적속성부분집합 선택 단계에서는 BPSO를 이용하여 후보속성부분집합에서 최적의 속성부분집합을 탐색한다. BPSO의 목적함수는 분류기의 정확도와 선택된 속성 수를 포함하는 다중목적함수(Multi-Object Function)을 이용하였다. 제안된 기법의 성능을 평가하기 위하여 유전자 데이터를 사용하였으며, 실험결과 기존의 방법들에 비해 우수한 성능을 보임을 알 수 있었다.

키워드 : 속성선택, 패턴분류, Binary Particle Swarm Optimization, 상호정보량

Abstract

In this paper, we proposed a feature selection method using Binary Particle Swarm Optimization(BPSO) and Mutual information. This proposed method consists of the feature selection part for selecting candidate feature subset by mutual information and the optimal feature selection part for choosing optimal feature subset by BPSO in the candidate feature subsets. In the candidate feature selection part, we computed the mutual information of all features, respectively and selected a candidate feature subset by the ranking of mutual information. In the optimal feature selection part, optimal feature subset can be found by BPSO in the candidate feature subset. In the BPSO process, we used multi-object function to optimize both accuracy of classifier and selected feature subset size. DNA expression dataset are used for estimating the performance of the proposed method. Experimental results show that this method can achieve better performance for pattern recognition problems than conventional ones.

Key Words : Feature selection, Pattern classification. Binary Particle Swarm Optimization, Mutual information.

1. 서 론

최근에는 고차원 데이터의 연산과 분류의 정확성을 높이기 위해서 다양하고 개선된 기술의 연구가 이루어지고 있다. 특히, 속성선택기법이 데이터 마이닝, 패턴 인식, 기계학습 등의 분야에서 크게 관심을 받고 있다. 속성선택기법은 원 데이터의 속성으로부터 좋은 성능에 영향을 주는 속성들만을 선택하거나 성능의 저하를 가져오는 속성들은 제거하여 분류기의 연산 효율성과 성능을 개선시키는 방법이다.

접수일자 : 2008년 11월 18일
완료일자 : 2009년 3월 25일
+ : 교신저자
본 연구는 보건복지가족부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임. (과제고유번호 : A040032)

속성선택기법의 단계는 크게 부분집합의 생성, 부분집합의 평가, 종료조건 판별, 선택된 속성부분집합의 유효성 평가 단계로 구분될 수 있다. 또한 부분집합의 생성 전략에 따라 전역적 탐색, 순차적 탐색, 무작위 탐색으로 구분된다. 부분집합의 평가를 위하여 다양한 방법들이 사용될 수 있는데 평가조건에 따라 필터(Filter) 방법, 랩퍼(Wrapper) 방법 그리고 혼합형(Hybrid) 방법으로 나눠질 수 있다[1]. 필터 방법은 분류기에 대해 독립적으로 속성부분집합을 평가하는 방법으로써 데이터들의 거리(distance) [2], 정보(information) [3], 의존성(dependence) [4], 일관성consistency [5] 등을 이용하여 선택된 속성부분집합을 평가한다. 필터방법에 의해 선택된 속성부분집합은 오직 원 데이터들의 특성만을 평가하고 속성들의 중요도나 우수성들을 판별하여 속성들을 선택한다. 즉, 데이터 속성들의 거리, 상호정보량, 중복도, 일관성 등을 평가하여 우수성의 순위 결정하고, 그 순위를 기반으로 사용자가 요구하는 속성 수만

큼 추출한다. 반면에 랩퍼 기법은 분류기의 성능에 의존적이다. 원 데이터에서 선택된 속성들을 분류기에 적용하고 선택된 속성부분집합이 우수한 성능을 보이는지 판별한다. 이 단계들을 설계자가 설정한 종료조건이 될 때 까지 반복적으로 연산한다. 이런 반복적 연산에 기인하여 필터 방법보다 연산수행시간이 많이 걸리는 단점이 있으나, 일반적인 성능은 필터 방법에 비해 우수하다. 혼합형 방법은 랩퍼 기법과 필터 기법을 융합한 구조로서 각각의 기법들의 단점을 보완하여 최적의 속성부분집합을 선택하는 방법으로서 최근에는 유전자 알고리즘과 같은 진화 알고리즘이나 퍼지 등을 융합한 속성선택 기법들이 많이 연구되어지고 있다[1][6].

한편, Particel Swarm Optimization(PSO) 기법은 생태계의 새,벌 등의 군집활동을 모방하여 최적화 알고리즘에 적용한 기법으로 여러 분야에 다양하게 적용되어져 왔다[7],[8]. 일반적인 PSO는 실수기반으로 속성선택문제에 적용하기에는 어려움이 있으며, 이를 해결하기 위해서 이진형태를 다룰 수 있는 Binary Particel Swarm Optimization (BPSO)가 속성선택문제에 많이 이용된다. 그러나 큰 속성을 가지는 데이터의 속성선택문제에는 BPSO만을 이용하는 것은 큰 메모리 공간과 연산에 비효율성 등의 단점이 존재하게 된다.

따라서 본 논문에서는 BPSO과 상호정보량을 적용한 개선된 속성선택기법과 BPSO의 효과적인 탐색을 위하여 적은 속성 수를 이용하여 최적의 분류기 성능을 만족하게 하는 목적함수를 제안하였다. 제안된 방법은 상호정보량을 이용한 후보 속성부분집합 선택과 BPSO를 이용한 최적 속성부분집합 선택 단계로 나누어진다. 제안된 방법의 성능을 평가하기 위하여 큰 속성을 가지는 유전자 데이터에 적용하였으며 그 유용성을 분석하였다.

2. 상호정보량과 PSO

2.1 상호정보량

속성선택문제에서 주요한 속성들은 출력에 대하여 중요한 정보들을 많이 포함하고 있고 반대로 그렇지 못한 속성들은 출력에 관해 적은 정보들을 포함한다. 분류 문제를 해결하기 위해서는 입력 속성에서 가능한 한 많은 정보들을 포함하도록 속성들을 선택 하여야한다. 이런 목적을 달성하기 위해 임의의 변수들의 정보를 측정하는 Shannon의 정보이론에서는 엔트로피(Entropy)와 상호정보량을 소개하였다. 기본적으로 엔트로피는 랜덤 변수들의 무질서도(uncertainty)를 측정하는 것이다. 만약 랜덤변수 X가 $p(x)=\Pr\{X=x\}, x\in\lambda$의 소스알파벳 λ을 가진다면 X의 엔트로피는 아래 식(1)과 같이 계산된다.

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

두 개의 랜덤 변수 X와 Y의 결합 엔트로피(joint entropy)는 아래 식으로 계산될 수 있다.

$$H(X,Y)=-\sum_{x\in\lambda}\sum_{y\in\delta}p(x,y)\log p(x,y) \tag{2}$$

위 식(2)에서 $p(x,y)$는 X와 Y의 결합확률밀도함수(joint probability density function)이다. 또한, Y를 알고 있을 때의 X의 조건부 엔트로피는 다음 식으로 계산될 수 있다.

$$H(X\quad Y)=-\sum_{x\in\lambda}\sum_{y\in\delta}p(x,y)\log p(y|x) \tag{3}$$

X와 Y의 공통의 정보량은 두 변수 사이의 상호정보량으로 정의 될 수 있고 아래 식으로 계산될 수 있다.

$$I(X;Y)=\sum_{x\in\lambda}\sum_{y\in\delta}p(x,y)\log\frac{p(x,y)}{p(x)\cdot p(y)} \tag{4}$$

수식 (4)에서, 두 변수 사이의 상호정보량 $I(X;Y)$ 크면 두 변수는 가까운 연관성을 가지고, 그렇지 않으면 두 변수는 연관성이 적거나 독립일 수 있다.

패턴인식 문제의 속성 선택에서 연속적인 속성을 F , 클래스를 C 라 정의하면 속성과 클래스간의 상호정보량은 다음과 같이 정의 된다.

$$I(F;C)=\sum_{f\in\lambda}\sum_{c\in y}p(f,c)\log\frac{p(f,c)}{p(f)\cdot p(c)} \tag{6}$$

상호정보량 $I(F;C)$이 클수록 속성 F가 클래스 C에 대해 많은 정보를 포함하는 것으로 정의되고 작으면 작을수록 속성 F가 클래스 C에 대해서 작은 정보를 포함하는 것으로 간주할 수 있다. 클래스에 대해 정보를 많이 포함한 속성을 선택함으로써 확률적으로 분류기의 성능을 우수하게 할 수 있고, 적은 정보량을 포함한 속성들을 제거함으로써 노이즈나 잘못된 데이터에 의한 분류기의 오차를 줄일 수 있다.

본 논문에서는 필터 기반의 전처리 선택으로서 이러한 상호정보량의 특징을 이용하여 각각의 속성들과 클래스 간의 상호정보량을 계산하고 순위별로 정리하여 각 속성들의 중요도를 평가하였다. 평가된 속성들을 기반으로 설정된 비율에 따라 하위 순위의 속성들을 제거하였다. 남은 속성들은 BPSO의 초기 particle 생성에 있어 후보속성부분집합으로 사용하였다.

일반적으로, 상호정보량 기반 속성선택은 순차적 탐색 방법의 하나인 탐욕 전방 선택(Greedy Forward Selection: GFS)으로 수행되어진다. 이 방법은 선택된 속성들의 빈 집합으로부터 시작하고 선택된 속성집합에 가장 큰 상호정보량을 포함한 입력속성을 하나씩 더해가면서 설정된 개수까지 반복한다. 상호정보량을 이용하는 일반적인 GFS는 아래와 같은 절차를 가진다[7].

[단계 1] 초기화(n개의 속성을 가지는 초기 집합 F, 공집합 S)

[단계 2] (클래스 C 를 이용하여 각각의 속성 F 에 대한 상호정보량 $I(C;F)$ 계산)

[단계 3] (상호정보량이 가장 큰 속성(f_i)을 선택하고 속성을 집합 S 에 저장)

[단계 4] (탐욕 전방선택) 설정된 속성 수만큼 반복하여 선택한다.

[단계 4-1] (결합 상호정보량(joint mutual information: $I(C;f_i,S)$)을 계산한다.)

[단계 4-2] (결합 상호정보량이 가장 큰 속성을 선택하고 선택된 속성집합 S에 추가한다.)

여기서, 결합 상호정보량 $I(C;f_i,S)$는 임의의 속성 f_i가 클래스에 대한 큰 상호정보량, 그리고 이미 선택된 속성들의 집합 S와의 상호정보량은 작은(이미 선택된 속성들과 중복도가 최소가 되는 것) 속성일수록 큰 값을 가지며 중요

도의 순위가 높게 평가된다. 그러나, 상호정보량을 이용한 이상적인 GFS의 단점은 결합상호정보량의 연산량이 속성의 수가 커지면 크게 증가한다는 점이다.

2.2 Particle Swarm Optimization

1995년에 진화형 계산기법의 일종으로 Particle Swarm Optimization(PSO)이 J.Kennedy와 R.Eberhart에 의해 제안되었다. PSO는 종래의 새나 물고기 무리의 움직임에 관한 연구로부터 유도되었다. 즉, 이러한 무리가 먹이를 찾아가는 과정에서 무리 전체가 정보를 공유한다는 가설과 무리 내부의 particle(무리 내의 각 개체를 지칭)이 지금까지 자기의 경험과 무리 전체에 공유되어 있는 정보를 기초로 하여 행동한다는 개념을 최적화 과정에 도입한 방법이라 할 수 있다.

이 방법은 다음과 같은 장점이 있다.[8]

(1) 알고리즘이 간단하고, 계산 시간이 짧으며 수렴성이 다른 진화연산에 비해 우수하다 .

(2) 연속형과 비 연속형의 문제 양쪽에 적용 가능하다.

PSO에서 각 particle은 지금까지의 탐색 중 최량의 목적함수 $F(pbest)$를 기억하고 있다. 또한 각 particle은 전체의 particle이 이제까지의 탐색과정에서 발견한 해 중에 최량의 해, 다시 말해 집단에서 발견한 해 중에 최량의 목적함수 값$F(gbest)$와 그 해의 위치 벡터의 정보를 공유한다. 각 particle은 현재의 위치 벡터와 속도벡터, 그리고 $pbest$,$gbest$를 이용해서 식(7) 의해 이동을 하게 된다. 또한 각 particle의 위치벡터의 수정은 현재의 위치와 수정된 속도를 이용해서 식(8)과 같이 행해진다.

$$v_k^{new} = w \cdot v_k^{old} + c_1 \cdot r_1 \cdot (pbest_k - x_k^{old}) + c_2 \cdot r_2 \cdot (gbest - x_k^{old}) \quad (7)$$

$$x_k^{next} = x_k^{now} + v_k^{new} \quad (8)$$

v_k^{old} : 현재 particle k의 속도벡터

x_k^{old} : 현재 particle k의 위치벡터

v_k^{new} : 수정된 particle k의 속도벡터

$pbest_k$: particle k가 현재까지 탐색 중 발견한 최량의 위치벡터

$gbest$: 전체 particle들이 지금까지 탐색 중 발견한 최량의 위치벡터

$r_{1,2}$: $U(0,1)$의 확률분포 값.

w, c_1, c_2: 가중치 계수

위의 수식을 이용한 PSO의 탐색절차는 아래와 같이 간략하게 설명될 수 있다.

[단계 1] 전체 에이전트에 대해 초기 위치 벡터 x_k와 속도벡터 v_k를 난수를 이용해 설정

[단계 2] 각각의 에이전트들이 지금까지 탐색하여 찾은 가장 우수한 값을 $pbest$, $pbest$들 중 가장 우수한 값을 $gbest$로 설정한다.

[단계 3] 식(7),식(8)를 이용하여 속도벡터 v_k, 위치벡터 x_k를 갱신한다.

[단계 4] 현재 에이전트의 위치에서 목적함수가 이전의 $pbest$, $gbest$보다 우수하면 그 값으로 각각을 갱신

[단계 5] 종료조건(반복횟수나 설정된 목적함수 값)이 만족하면 종료, 그렇지 않으면 단계 3에서 반복.

3. 제안된 속성선택 기법

본 논문에서는 속성선택문제에서 필터형태의 방법과 랩퍼형태의 기법들의 단점을 상호보완할 수 있는 융합된 구조의 개선된 속성선택기법을 제안하였다. 필터형태의 속성선택 기법은 고차원 데이터들에 대해 랩퍼기법에 비해 연산량이 적은 반면 성능이 랩퍼기법에 비해 낮고, 반대로 랩퍼형태의 속성선택 기법들은 성능이 우수한 반면 연산량이 증가하는 단점을 가지고 있다.

제안된 알고리즘은 앞서 언급한 단점들을 극복하기 위해서 크게 두 부분의 속성선택 단계로 이루어진다. 첫째로 상호정보량을 이용한 후보 속성부분집합 선택 단계에서는 클래스와 각 속성들의 상호정보량만을 이용하여 후보속성부분집합을 생성하고, 두 번째 단계 BPSO를 이용한 최적속성부분집합에서는 전 단계에서 선택된 후보속성부분집합을 이용하여 최적의 속성부분집합을 선택한다. 그림 2는 제안된 알고리즘의 순서를 나타낸다.

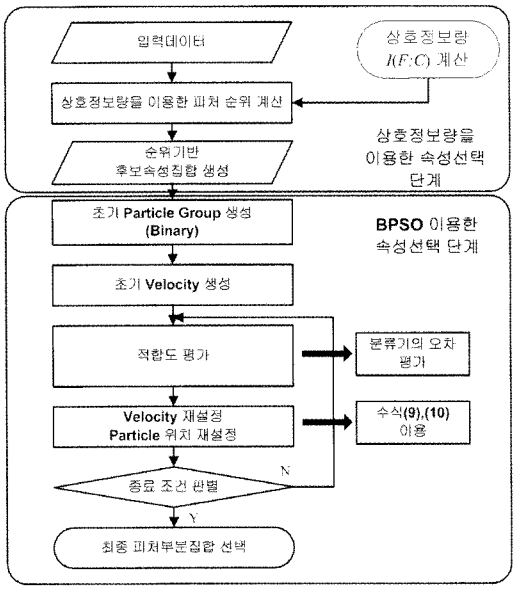

그림 2. 제안된 기법의 순서도

Fig. 2. Flowchart of proposed method.

3.1 상호정보량을 이용한 후보 속성부분집합 선택

상호정보량을 이용한 속성선택기법에서 앞에서 설명한 결합상호정보량은 속성의 수가 큰 데이터를 처리할 때 연산량의 증가와 설계자가 선택하는 목표 속성 수에 따라 성능의 큰 차이를 보이는 단점을 가지고 있다. 본 논문에서는 결합상호정보량의 큰 연산량을 줄이기 위해서 클래스 C 와 속성 F의 상호정보량$I(C;F)$ 만을 이용한다. 이 방법은 속성들과 클래스간, 속성과 속성간의 상호정보량을 이용하는 방법보다 각각의 속성의 중요도를 평가하는데 큰 불확실성이 존재할 수 있다. 그러나 본 논문에서는 상호정보량을 이용하여 최적 속성부분집합을 평가하는 것이 아니라 오직 후보 해만을 선택하는데 사용되어지기 때문에 속성들에 대한 정확한 상호정보량의 평가는 불필요하다.

아래는 본 논문에서 상호정보량이 이용되는 부분의 절차를 나타내었다.

[단계 1] 각각의 속성F와 클래스 C 간의 상호정보량 $I(C;F)$을 계산한다.

[단계 2] 계산된 상호정보량을 기반으로 상호정보량이 높은 순으로 정렬한다.

[단계 3] 설계자에 의해 설정된 후보속성 수만큼을 순위별로 선택하여 후보속성부분집합을 생성한다.

3.2 BPSO를 이용한 최적 속성부분집합 선택

일반적인 PSO 알고리즘은 실수기반으로 최적화 문제에 적용되어져왔다. 그러나 다양한 최적화 문제에서 실수표현보다 이진표현으로 처리되었을 때 우수한 성능을 보이는 문제들도 많이 존재한다. 이런 이진 표현들을 위해서 Kennedy와 Eberhart 는 이진 표현이 가능한 Binary Particle Swarm(BPSO) 를 제안하였다[9]. 본 논문에서는 속성들의 선택문제를 각각의 particle들에 대해 선택되었을 때는 '1', 선택되지 않을 때는 '0'으로 표현되는 BPSO를 이용하여 속성부분집합을 생성하였다. BPSO는 이진공간에서 문제를 표현하고 탐색하기 위하여 아래의 속도,위치 갱신 방정식을 사용한다.

$$v_{pd}^{new} = w \cdot v_{pd}^{old} + c_1 \cdot r_1 \cdot (pbest_{pd} - x_{pd}^{old}) + c_2 \cdot r_2 \cdot (gbest_{pd} - x_{pd}^{old}) \quad (p = 1,2,3 \cdots n,\ d = 1,2,3 \cdots m) \tag{9}$$

$$S(v_{pd}^{new}) = \frac{1}{1+e^{-v_{pd}^{new}}},\ x_{pd}^{\neq w} = \begin{cases} 0 & \text{if } r \geq S(v_{pd}^{new}) \\ 1 & \text{if } r < S(v_{pd}^{new}) \end{cases} \tag{10}$$

$$V_{\max} = 6,\ V_{\min} = -6$$

v_{pd}^{old} : 현재 particle p의 속도벡터

x_{pd}^{old} : 현재 particle p의 위치벡터

v_{pd}^{new} : 수정된 particle p의 속도벡터

pd : particle p의 d번째 파라미터

n : particle 의 총수 , m : 속성의 총 수

w, c_1, c_2: 가중치 계수

$pbest_p$: particle p가 현재까지 탐색 중 발견한 최량의 위치벡터

$gbest_p$: 전체 particle들이 지금까지 탐색 중 발견한 최량의 위치벡터

$V_{\max}, V_{\min}$: 속도벡터의 최대, 최소 범위.

그림 3은 BPSO에서 각각의 particle들이 속성부분집합을 표현하는 방법을 보였다. particle k 는 원 데이터 속성에서 [7,5,6,2] 번째의 속성들만을 선택해서 속성부분집합을 생성한다. 이런 방법으로 각각의 particle들에서 속성부분집합을 생성하게 되고 목적함수에 의해 평가를 한다.

속성선택을 위한 목적함수는 속성 수를 최대한 적게 이용하여 분류기의 오차가 최소로 하는 다목적함수를 사용하며 두 조건들의 반영 비율을 조절하는 가중치를 이용하는 방법이 일반적이다[10][11].

$$F_1 = w \cdot acc(x) + (1-w) \cdot (1/s(x)) \tag{11}$$

$$F_2 = w \cdot acc(x) - (1-w) \cdot \frac{s(x)}{totalfeat} \tag{12}$$

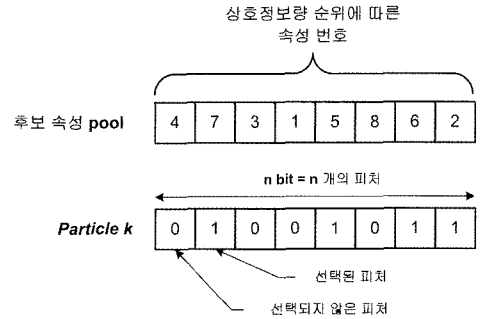

그림 3. BPSO의 particle에서의 속성들의 표현

Fig. 3. Expression of features in a particle of BPSO

위 식에서 w는 분류기의 성능과 선택된 속성사이의 가중치(0~1), $acc(x)$는 분류기의 성능(분류기의 정확도 : 0~1), $s(x)$는 선택된 속성들의 수, $totalfeat$는 전체 속성의 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. 그러나 식(11)과 식(12)의 두 번째 항에서 원 데이터의 속성($totalfeat$)이 크거나 선택된 속성($s(x)$)이 크면 두 번째 항의 값이 아주 작아지기 때문에 탐색성능에 큰 문제를 가져올 수 있다. 예를 들어 w=0.5, 분류기의 정확도가 0.5 , 선택된 속성수가 30개, 총 속성수가 1000개일 때, 식(11)의 두 번째 항은 0.015, 식(12)의 두 번째 항은 0.015로 되어 첫째항 0.25에 비해 너무 작은 값을 가지기 때문에 첫째항 값에 크게 의존적이게 된다. 따라서 본 논문에서는 이러한 단점을 극복하기 위해서 아래 식(13)을 제안하였다.

$$F_3 = w \cdot acc(x) + (1-w) \cdot \left(1 - \frac{S\_fea}{A\_fea}\right) \tag{13}$$

S_fea : 선택된 속성의 수, A_fea : 전체 속성의 수

식(13)에서 위의 예를 적용해 보면 약 0.49로 되어 동일한 소수점 자리에서 값을 가지게 된다. 또한, 분류기의 정확도가 높아질수록 선택되는 속성 수와 비슷한 비중으로 목적함수 값을 탐색하게 되어 두 조건을 충족하는 값으로 수렴이 가능하다.

제안된 알고리즘의 BPSO 알고리즘의 속성선택 순서는 다음과 같다.

[단계 1] 상호정보량에 의해 생성된 후보속성부분집합을 이용하여 초기 particle들을 생성한다.

[단계 2] 초기 속도 벡터 v^{init}를 랜덤함수를 이용하여 설정한다.

[단계 3] 생성된 초기 particle들을 분류기에 적용하여 각각의 성능을 식(13)을 이용하여 평가한다. 평가된 성능을 기반으로 지금까지 particle k가 찾은 가장 우수한 위치를 $pbest_k$로 설정하고 모든 particle들이 지금까지 찾은 가장 우수한 위치를 $gbest$ 설정한다.

[단계 4] 수식(9)와 수식(10)을 이용하여 새로운 particle의 위치와 속도를 업데이트 한다.

[단계 5] 종료 조건(설정된 반복횟수, 목적함수의 최대값)을 판별하여 조건을 만족할 때까지 단계 3~5를 반복한다.

4. 시뮬레이션 및 결과 고찰

제안된 속성선택 알고리즘을 평가하기 위하여 표 1의 3개의 유전자 발현 데이터[3]에 적용하고 기존의 속성선택 알고리즘과 비교 분석하였다. 실험을 위하여 필터기법에서 선택되는 속성 수는 200개로 설정하였으며, k-NN알고리즘에 k=1로 설정하였다. BPSO 알고리즘에서는 반복횟수는 20회, w, c_1, c_2는 각각 0.5, 2, 2 로 설정하였다. 각각의 실험들은 결과의 타당성을 위해서 Leave-One-Out Cross-Validation (LOOCV) 방법을 이용하였다. LOOCV 방법은 데이터의 수가 적을 때 실험결과의 타당성을 확보하기 위한 방법으로 순서대로 한 개의 데이터를 확인데이터로 나머지 데이터들을 학습데이터로 하여 모든 데이터들이 한 번은 확인데이터가 될 때까지 반복하여 실험데이터의 수만큼의 결과 값들을 얻는다. 얻어진 결과 값들을 평균을 취하여 최종 출력 값으로 평가한다.

MIFS(Mutual information based feature selection) 방법은 상호정보량만을 이용하여 50개의 속성을 이용한 방법 중 가장 우수한 결과이며, GA-MI방법은 유전자 알고리즘과 상호정보량을 이용한 속성선택기법으로 집단크기는 20, 교배확률 0.7, 돌연변이 확률 0.1, 반복횟수를 20, 이점교배 방법과 단순 돌연변이를 이용하여 수행하였다.

표 1. 실험에서 사용된 유전자 발현 데이터.
Table 1. gene expression dataset used in our experiment.

데이터 명	속성 수	데이터 수	검증데이터
LEUKEMIA	7070	72	2
LYMPHOMA	4026	96	9
COLON CANCER	2000	62	2

그림 4-6은 각각의 데이터들에 대해서 3장의 식 (11),(12),(13)의 목적함수 F_1, F_2, F_3에 대해 BPSO의 최적의 속성 수 탐색결과를 보였다. LOOCV의 특성상 첫 번째 데이터가 확인데이터(Test data), 나머지 데이터가 학습데이터(Training data) 일 때만 그림에 나타내었다. 그림 4에서 보이듯이 Lekumia 데이터에 대해서 목적함수 F_1, F_2 는 약 20세대에서 더 이상 속성 수를 반영하여 탐색하지 못하고 지역해(Local optimal feature subset)로 수렴하는 것을 알 수 있다.

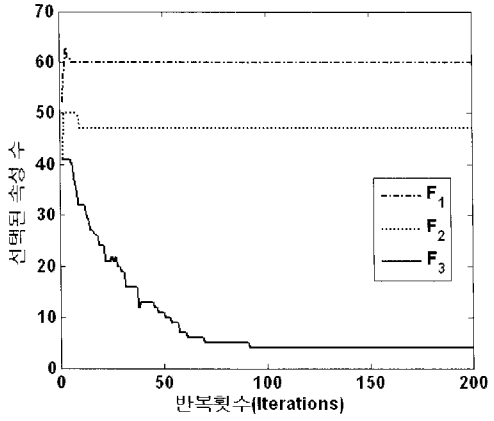

그림 4. 최적속성 부분집합 탐색과정에서 선택된 속성(Lekumia data) .
Fig. 4. The optimal number of selected feature in each generation.

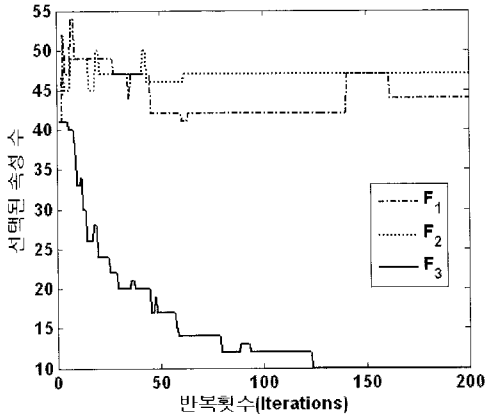

그림 5. 최적속성 부분집합 탐색과정에서 선택된 속성(Lymphoma data).
Fig. 5. The optimal number of selected feature in each generation.

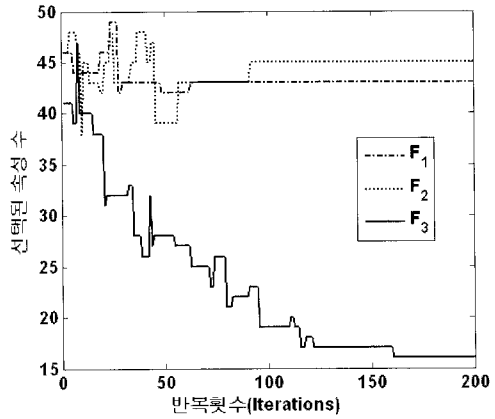

그림 6. 최적속성 부분집합 탐색과정에서 선택된 속성(Conlon Cancer data).
Fig. 6. The optimal number of selected feature in each generation.

그러나 제안된 목적함수 F_3은 세대의 증가에 따라 속성 수를 잘 반영하여 탐색하는 것을 알 수 있다. 그림 5와 그림 6의 두 데이터에 대한 속성 수를 반영한 탐색과정도 비슷한 경향을 보이고 있다. 그림 7은 제안된 방법으로 선택된 최적 속성부분집합을 이용한 각 세대별 분류기 오차를 나타내었다. 그림 7에서 보이듯이 제안된 알고리즘이 분류기의 오차를 최소로 하는 속성부분집합들을 잘 탐색하는 것을 알 수 있다. 표 2에서는 목적함수 F_3을 이용하여 제안된 알고리즘과 다른 속성부분집합 방법들의 성능을 비교하였다. 표 2에서 알 수 있듯이 제안된 알고리즘 다른 알고리즘에 비해서 일반적으로 우수한 성능을 보임을 알 수 있다.

5. 결 론

본 논문에서는 이진 PSO와 상호정보량을 이용한 속성선택 기법을 제안하였다. 제안된 방법은 크게 상호정보량을 이용한 후보 속성부분집합을 선택하는 단계와 BPSO를 이

용한 최적 속성부분집합을 선택하는 단계로 나누어진다. 또한, BPSO의 탐색성능을 개선하기 위하여 다목적함수(Multi-object function)을 사용하였다. 제안된 방법은 속성과 클래스간의 상호정보량만을 이용하여 연산의 효율성을 개선하였으며, BPSO를 이용하여 후보속성부분집합에 대한 최적의 속성부분집합을 선택함으로써 지역해(Local optimal feature subset)를 피하는 장점을 가지고 있다. 제안된 방법을 유전자 발현데이터에 적용하고 LOOCV를 이용하여 실험의 타당성을 확보하였다. 실험결과 일반적으로 다른 방법에 비해 우수한 성능을 보임을 알 수 있었으며 향후 다양한 데이터로의 확장적용이 가능할 것으로 예상한다.

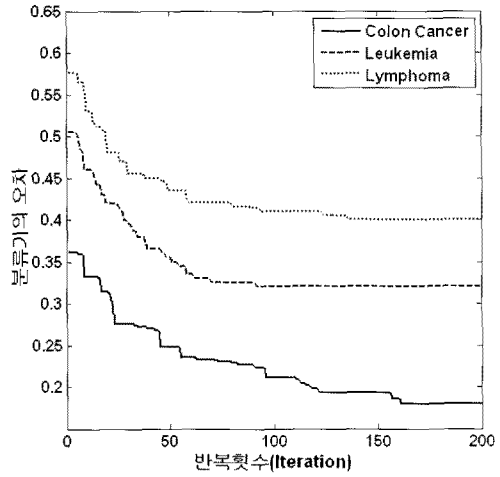

그림 7. 각 데이터에 대한 최적 속성 부분집합 탐색과정.
Fig. 7. Search process of optimal feature subset for each dataset.

표 2. 제안된 방법과 기존방법들의 성능비교
Table 2. comparition of the proposed method and others

Dataset	MIFS		GA-MI		제안된 방법 (BPSO+k-NN)	
	속성 수	오차	속성 수	오차	속성 수	오차
LEUKEMIA	19	27.39±2.1	19±1.8	4.16±2.2	21±1.1	3.82±1.7
LYMPHOMA	21	28.11±5.4	18±2.5	5.33±1.9	17±2.5	4.62±1.2
COLON CANCER	31	25.88±2.8	24±4.3	1.84±2.4	19±2.7	1.54±2.2

참 고 문 헌

[1] H. Liu, L. Yu, "Toward Integrating Feature Selection Algorthms for Classification and Clustering," *IEEE Trans. Knowledge and Data Engineering.*, vol. 17, No.4, pp. 491-502, 2005.

[2] H. Almuallim and T.G. Dietterich, "Learning with Many Irrelevant, Features," *Proc. Ninth Nat'l conf. Artificial Intellignece*, vol.69, no.1-2, pp. 279-305, 1994.

[3] H. Peng, F. Long and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy", *IEEE Trans.Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, August 2005.

[4] M.A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning," *Proc. 17th Int'l conf. Machine Learning*, pp. 359-366, 2000.

[5] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining. Boston* : Kluwer Academic. 1998.

[6] 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.

[7] R. Battit., "Using mutual information for selecting features in supervised neural net learning", *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537-550, 1994.

[8] J. Kennedy, R.C. Eberhart, "Particle Swarm Opimization" *IEEE Int'l conf. Neural Networks*, vol. 4, pp.1942-1948, 1995.

[9] J. Kennedy, R. Eberhart,, "A discrete binary version of the particle swarm algorith.", *IEEE interna. Conf. Computational Cybernetics and Simulation*, vol. 5, pp. 4104-4108,1997.

[10] L.Y. Chuang, H.W. Chang, C.J. Tu, C.H. Yang, "Improved Binary PSO for feature selection using gene expression data" *computational Biology and Chemistry*, vol. 32, no.1, pp. 29-38, 2008.

[11] 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.

저 자 소 개

조재훈(Jae Hoon Cho)
2008년 2월 제 18권 제 1호 참조

이대종(Dae Jong Lee)
2008년 2월 제 18권 제 1호 참조

송창규(Chang Kyu Song)
2008년 2월 제 18권 제 1호 참조

전명근(Myung Geun Chun)
2008년 2월 제 18권 제 1호 참조