DOI: 10.3745/KIPSTB.2008.15-B.4.331

생태계 모방 알고리즘 기반 특징 선택 방법의 성능 개선 방안

윤 철 민[†] · 양 지 훈[††]

요 약

특징 선택은 기계 학습에서 분류의 성능을 높이기 위해 사용되는 방법이다. 여러 방법들이 개발되고 사용되어 오고 있으나, 전체 데이터에서 최적화된 특징 부분집합을 구성하는 문제는 여전히 어려운 문제로 남아있다. 생태계 모방 알고리즘은 생물체들의 행동 원리 등을 기반으로 하여 만들어진 진화적 알고리즘으로, 최적화된 해를 찾는 문제에서 매우 유용하게 사용되는 방법이다. 특징 선택 문제에서도 생태계 모방 알고리즘을 이용한 해결방법들이 제시되어 오고 있으며, 이에 본 논문에서는 생태계 모방 알고리즘을 이용한 특징 선택 방법을 개선하는 방안을 제시한다. 이를 위해 잘 알려진 생태계 모방 알고리즘인 유전자 알고리즘(GA)과 파티클 집단 최적화 알고리즘(PSO)을 이용하여 데이터에서 가장 분류 성능이 우수한 특징 부분집합을 만들어 내도록 하고, 최종적으로 개별 특징의 사전 중요도를 설정하여 생태계 모방 알고리즘을 개선하는 방법을 제안하였다. 이를 위해 개별 특징의 우수도를 구할 수 있는 mRMR이라는 방법을 이용하였다. 이렇게 설정한 사전 중요도를 이용하여 GA와 PSO의 진화 연산을 수정하였다. 데이터를 이용한 실험을 통하여 제안한 방법들의 성능을 검증하였다. GA와 PSO를 이용한 특징 선택 방법은 그 분류 정확도에 있어서 뛰어난 성능을 보여주었다. 그리고 최종적으로 제시한 사전 중요도를 이용해 개선된 방법은 그 진화 속도와 분류 정확도 면에서 기존의 GA와 PSO 방법보다 더 나아진 성능을 보여주는 것을 확인하였다.

키워드 : 기계학습, 특징 선택, 생태계 모방 알고리즘, 진화적 알고리즘

Performance Improvement of Feature Selection Methods based on Bio-Inspired Algorithms

Chulmin Yun[†] · Jihoon Yang[††]

ABSTRACT

Feature Selection is one of methods to improve the classification accuracy of data in the field of machine learning. Many feature selection algorithms have been proposed and discussed for years. However, the problem of finding the optimal feature subset from full data still remains to be a difficult problem. Bio-inspired algorithms are well-known evolutionary algorithms based on the principles of behavior of organisms, and very useful methods to find the optimal solution in optimization problems. Bio-inspired algorithms are also used in the field of feature selection problems. So in this paper we proposed new improved bio-inspired algorithms for feature selection. We used well-known bio-inspired algorithms, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), to find the optimal subset of features that shows the best performance in classification accuracy. In addition, we modified the bio-inspired algorithms considering the prior importance (prior relevance) of each feature. We chose the mRMR method, which can measure the goodness of single feature, to set the prior importance of each feature. We modified the evolution operators of GA and PSO by using the prior importance of each feature. We verified the performance of the proposed methods by experiment with datasets. Feature selection methods using GA and PSO produced better performances in terms of the classification accuracy. The modified method with the prior importance demonstrated improved performances in terms of the evolution speed and the classification accuracy.

Key Words : Machine Learning, Feture Selection, Bio-Inspired Algorithm, Evolutionary Algorithms

1. 서 론

기계 학습 (Machine Learning)의 분야에서, 분류기의 분류 성능을 향상시키는 것은 가장 중요한 목표 중 하나이다. 이렇게 분류의 성능을 향상시키기 위해서는 여러 가지 보조적인 방법들이 사용되는데, 특징 선택 (Feature Selection)도 그 중 하나의 방법으로, 일반적으로 주어진 데이터에서 분류기의 분류 목적에 밀접하게 연관되어 있는 특징들만을 골라내어 새로운 데이터의 집합을 만들어 내는 것으로 정의된다[1,2].

특징 선택 방법은 그 중요성이 부각된 이후 현재까지 다양한 방법들이 연구되어 왔다. 여러 가지 효율적인 방법들

† 정 회 원 : (주)다이퀘스트 선임연구원
†† 정 회 원 : 서강대학교 컴퓨터학과 부교수
논문접수 : 2008년 1월 4일
수 정 일 : 1차 2008년 3월 10일, 2차 2008년 4월 30일, 3차 2008년 5월 14일
심사완료 : 2008년 5월 17일

이 제시되어 오고 또 사용되어 오고 있으나, 아직까지도 더 나은 성능을 내기 위해 개선해야 할 점들이 남아 있다.

이러한 가운데 근래에 제시되어 오는 효율적인 특징 선택 방법의 하나로 생태계 모방 알고리즘(Bio-inspired Algorithm)을 사용하는 방법이 있다. 생태계 모방 알고리즘은 생태계의 생물체들의 행동 습성을 기반으로 한 알고리즘으로, 특히 다양한 해가 존재하는 공간에서 주어진 문제의 최적 해(Optimal Solution)를 찾아내는 데에 효과적인 알고리즘으로 알려져 있다. 이런 이유로 특징 선택 문제에서 생태계 모방 알고리즘은 유용한 해결방법 중 하나로 제시되어 왔다.

본 논문에서는 이러한 생태계 모방 알고리즘을 이용한 특징 선택 방법을 수정 보완하는 방법에 대해 논의한다. 기본적인 생태계 모방 알고리즘 기반 특징 선택을 기반으로 하여 특징의 사전 중요도(Prior Importance)의 개념을 생태계 모방 알고리즘에 결합해 기존 알고리즘을 수정하였다. 그리고 수정된 알고리즘을 통해 기존의 기본적인 생태계 모방 알고리즘을 사용하였을 때 보다 더 나은 특징 선택 성능을 이끌어 낼 수 있음을 데이터를 사용한 실험을 통해 제시한다.

2. 특징 선택의 개념과 특징

일반적으로 특징 선택 문제는 다음과 같이 정의된다.

어떤 학습(분류) 시스템 하에서 원본 데이터가 주어졌을 때, 가장 좋은 성능을 보여줄 수 있는 데이터의 부분집합(Subset)을 원본 데이터로부터 찾아내는 것[3]

이러한 특징 선택 과정을 통해 우리가 얻을 수 있는 이점은 크게 두 가지로 생각할 수 있다. 첫 번째로, 원본 데이터에 비해 줄어든 크기의 데이터를 얻을 수 있다는 것이다. 이에 따라 원본 데이터를 사용하였을 때에 비해서 더 빠른 시간에 연산을 마칠 수 있다[2].

특징 부분집합을 생성하는 과정은 주어진 원본 데이터에서 구성 가능한 특징의 부분집합, 즉 후보 집합군(Candidate Sets)을 탐색하는 일로 볼 수 있다. 이 과정에서 우리는 크게 두 가지 측면을 고려하여야 한다. 첫 번째는 탐색의 시작점(Starting Point)을 정하는 것이고, 두 번째는 탐색 전략(Search Strategy)을 생각하는 것이다[1,2]. 일반적으로 아무 특징도 선택되지 않은 상태에서 부분집합의 크기를 키워가는 방향으로 탐색을 진행하는 '순차적 선택(Forward Selectioin)' 방법과, 반대로 모든 특징이 선택되어 있는 상태에서 시작하여 특징의 부분집합의 크기를 줄여나가는 '역행적 제거(Backward Elimination)' 방법을 생각할 수 있다[1]. 이 외에도 양방향 방법, 무작위 지점 시작 방법 등이 있다. 또한 탐색 전략에 있어서는, 후보 집합의 완전한 탐색은 N개의 특징이 있다고 할 때 2^N번의 탐색과정을 거치게 되므로, 특징 수가 많은 데이터에서는 시간의 제약을 생각하여 여러 탐욕적 탐색이나 언덕 오름 탐색 등 효율적인 탐색 방법을 생각해 보아야 한다.

생성된 특징 부분집합의 유용성 평가에는 크게 두가지 방법이 있다. 첫 번째 방법은 독립적인 특징 부분집합 평가 기준을 두는 것으로, 필터 방법 (Filter Method) 이라고도 부른다. 선택된 특징 부분집합의 우수성을 부분집합 안의 특징들과 분류 기준 사이의 고유한 속성을 이용하여 평가하는 방법이다[1,2].상호 정보량 측정(Mutual Information Measurement), 정보 획득량 측정(Information Gain Measurement), 카이-제곱(Chi-Square) 측정 등의 방법들이 필터 방법의 특징 부분집합 평가 방법으로 이용되고 있다[4]. 두 번째 방법은 첫 번째와는 반대로 종속적인 특징 부분집합 평가 기준을 사용하는 것으로, 래퍼 방법(Wrapper Method)으로도 알려져 있다. 이 방법은 특정한 평가 기준을 두지 않고 직접적으로 분류기를 사용하여 그 특징 부분집합의 성능을 평가하는 것으로, 사전에 어떤 분류기를 사용하여 데이터의 분류를 행할 것인지를 결정하고, 매번 특징 부분집합이 생성될 때마다 해당 분류기를 사용해 데이터를 분류하여 그 분류 성능을 해당 특징 부분집합의 우수성 척도로 삼는다.

3. 생태계 모방 알고리즘 기반의 특징 선택 방법

3.1 생태계 모방 알고리즘을 이용한 특징 선택

2장에서 설명한 내용들을 토대로 유용한 특징 선택 방법들이 개발되고 사용되어 오고 있으나, 그 방법적인 면에서 몇 가지 한계를 지적할 수 있다. 그 중 하나는 주로 필터 방법에서 제시되는 문제로, 설정한 특정 평가 기준이 분류기의 정확도와는 직접적인 연관성이 떨어질 수 있다는 것이다. 이 문제는 직접 분류기로 특징의 평가 기준을 삼는 래퍼 방법으로 인해 보완할 수 있으나, 수행 속도 면에서는 필터 방법에 비해 뒤처진다는 단점이 있다. 또 하나의 문제는 특징 부분집합을 구성하는 데에서 오는 문제이다. 개개의 특징에 대한 우수성은 알 수 있으나, 그 특징들을 어떻게 조합하였을 때 가장 좋은 성능을 보이는 집합을 구성할 수 있는지에 대해서는 쉽게 생각하기 힘들다. 다른 특징들과 조합되어 좋은 성능을 보일 수 있는 특징들이 존재하기 때문에, 특징들의 최적 조합을 찾아내는 것은 매우 어려운 문제이다.

이러한 기존 특징 선택 방법의 한계를 극복하기 위해 제시된 방법들 중 하나가 생태계 모방 알고리즘을 사용한 방법이다. 생태계 모방 알고리즘은 기본적으로 어느 주어진 생태계 안에서 각각 하나씩의 가능한 해 (Solution)를 갖는 개체들이 모인 개체군이 각 알고리즘 나름의 진화연산을 수행하면서 최적의 (Optimal) 해 집단을 형성해 가는 것을 주요 목적으로 한다[5].

이러한 생태계 모방 알고리즘은 특히 최적화 문제에 있어 유용한 해결 방법 중 하나로 잘 알려져 있다. 대표적인 생태계 모방 알고리즘으로는 유전자 알고리즘(Genetic Algorithm, GA), 개미 군집 최적화 알고리즘(Ant Colony Optimization, ACO), 파티클 집단 최적화 알고리즘(Particle Swarm Optimization, PSO) 등이 있으며, 이 알고리즘들을 사용한 특징 선택 방법도 여러 연구를 통해 제시되어 오고 있다[6-14].

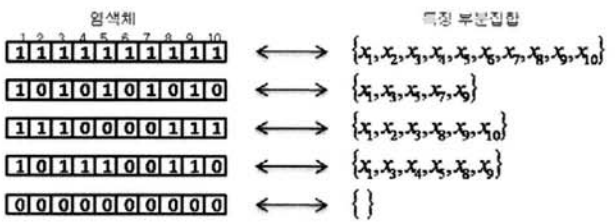
본 논문에서는 특징 선택 방법에 사용할 기본 생태계 모방 알고리즘으로 유전자 알고리즘과 파티클 집단 최적화 알고리즘의 두 가지 방법을 사용할 것이다. 본 논문에서 사용하는 방법은 두 알고리즘을 이용한 래퍼 방식의 특징 선택으로서, 생태계 모방 알고리즘의 방식을 통하여 최적의 특징 집합을 찾아내는 문제를 해결할 수 있고, 여기에 래퍼 방식을 사용하여 분류기의 정확도와 직접 연관된 특징 집합을 찾아낼 수 있어 기존 방법들보다 더 나은 특징 선택 성능을 보일 수 있다.

3.2 유전자 알고리즘 (GA)을 이용한 특징 선택

유전자 알고리즘(GA)은 널리 알려진 생태계 모방 알고리즘의 하나로, 생명체의 유전과 진화의 과정을 인공적으로 모사 하여 최적의 해(Solution)를 찾는 알고리즘이다[5,6]. 주로 염색체와 같은 자료구조를 사용하여 각 염색체가 하나의 임의의 가능한 해를 갖고 확률적으로 개체의 일부 혹은 전체가 선택(Selection), 교배(Crossover), 돌연변이(Mutaion)의 유전자 연산자(Genetic Operator)에 의해 다음 세대로 유전되면서 점점 해 집단이 최적의 해로 수렴해간다[15].

GA를 어떤 해결하고자 하는 문제에 적용할 때 가장 중요하게 고려하여야 하는 문제는 크게 두 가지로, 염색체 자료구조 안에서의 적절한 해의 표현(Representation) 문제와 적합도 함수(Fitness Function)를 설정해 주는 문제이다[1,15].

본 논문에서 사용한 방법은 다음과 같다. 해의 표현 방법, 즉 염색체의 구성 방법을 살펴보면, 하나의 염색체는 데이터의 특징 수만큼의 길이를 가지는 비트 문자열(Bit String)로 구성되어 각각의 비트는 0 또는 1의 값을 가진다. 각 비트는 데이터 안에서 해당하는 각 특징을 나타낸다. 즉 문자열의 i번째 비트는 데이터에서 i번째 특징을 나타내며, 0인 경우에는 해당 특징이 그 특징 부분집합 속에 포함되어 있지 않음을, 1인 경우는 포함되어 있음을 나타낸다. 이와 같은 방식을 사용해 하나의 염색체는 하나의 가능한 특징 부분집합을 나타내도록 한다. (그림 1)에서 이 과정을 도식화하여 나타내었다.



(그림 1) 특징 부분집합을 비트 문자열로 표현하는 예

적합도 함수는 각 염색체, 즉 만들어진 해들이 어느 정도나 최적 해에 근접한 해인지를 판별하는 기준이 되는 함수이다. 본 특징 선택 방법에서는 염색체가 표현하는 특징 부분집합을 직접 분류기에 적용하여 그 분류 성능을 측정하는 것을 적합도 함수로 삼았다. 즉 이 방법은 매번 만들어지는 특징 부분집합을 직접 분류기에 적용하여 그 우수성을 평가한다는 점에서 2장에서 설명한 래퍼 방법을 전형적으로 따른다고 볼 수 있다.

3.3 파티클 집단 최적화 (PSO)를 이용한 특징 선택

파티클 집단 최적화 알고리즘(PSO)은 GA와 함께 잘 알려진 또 하나의 생태계 모방 알고리즘 중 하나로서, 새, 물고기, 벌 등 군집 생활을 하는 동물들의 행동 습성을 모방하여 최적의 해를 찾는 알고리즘이다[16-18]. 이 방법은 여러 개의 파티클 들이 해 공간(Solution Space) 안에서 흩어져 있어, 반복을 거듭하며 좀 더 나은 해에 가까운 위치로 자신들의 위치를 변화시켜 가며 점점 파티클 집단이 최적의 해를 찾아가는 방향으로 수렴하게 된다.

마찬가지로 PSO를 해결하고자 하는 문제에 적용할 때 가장 중요하게 고려하여야 하는 것은 해를 각 파티클의 위치벡터로 표현하는 문제와 적합도 함수의 설정이다. 본 논문에서는 GA의 경우와 유사한 방식을 사용하였다. 먼저 파티클의 위치벡터 구성은 GA에서의 염색체 구성과 동일하게 특징 수만큼의 길이를 갖는 이진 문자열로 구성하였고, 각 특징에의 대응 역시 GA에서와 동일한 방식을 따른다. 또한 적합도 함수의 설정 역시 유전자 알고리즘과 마찬가지로 각 파티클의 위치벡터가 나타내는 특징 부분집합을 분류기에 적용하여 그 분류 성능을 측정하는 방식으로 설정하였다.

이 때, 각 파티클의 위치 벡터를 매 반복마다 갱신하는 방법은 Kennedy와 Eberhart에 의해 제시된 이진 문자열을 이용한 PSO 연산 구현[16]을 그대로 적용하였다. D-차원(D-Dimension)의 해 공간에서 N개의 파티클들이 매번 반복(Iteration) 을 거듭하며 움직인다고 할 때, t번째 반복에서의 i 번째 ($1 \le i \le N$) 파티클의 위치(Position)는 벡터형으로 $X_i^{(t)} = (x_{i1}, x_{i2}, ..., x_{iD})$와 같이 표현된다. 반복을 거치며 각 파티클들은 위치를 변화시키게 되는데, 두 개의 참고점을 고려하게 된다. 하나는 각각의 파티클들이 가장 우수성이 좋았을 때의 위치 벡터(Personal Best Position, pbest)이고 다른 하나는 전체 파티클들을 통틀어서 가장 우수성이 좋았을 때의 위치 벡터(Global Best Position, gbest)이다. i번째 파티클의 pbest P_i와 전체 파티클 안의 gbest P_g는 $P_i = (p_{i1}, p_{i2}, ..., p_{iD})$, $P_g = (p_{g1}, p_{g2}, ... p_{gD})$와 같이 표현되고, t 번째의 반복 때마다 파티클의 위치 벡터를 변화시켜주는 요소인 속도 (Velocity) 벡터 $V_i^{(t)}$는 $V_i^{(t)} = (v_{i1}, v_{i2}, ..., v_{iD})$와 같이 표현된다. X_i 의 d번째 값인 x_{id} 는 0 또는 1의 값을 가지게 되고, 매 반복마다 다음과 같은 식을 통해 값이 역시 0 또는 1로 변화된다.

$$v_{id}^{(t)} = v_{id}^{(t-1)} + c_1 (p_{id} - x_{id}^{(t-1)}) + c_2 (p_{gd} - x_{id}^{(t-1)})$$

$$if\ \rho_{id} < s(v_{id}(t))\ ,\ then\ x_{id}(t) = 1; else\ x_{id}(t) = 0$$

$$s(v_{id}) = \frac{1}{1 + \exp(-v_{id})}$$

p_{id}, p_{gd}, v_{id}는 각각 P_i, P_g, V_i의 d번째 값을 의미한다. 그리고 $s(v_{id})$는 v_{id}값의 크기를 [0.0, 1.0] 사이의 값으로 조

절해주기 위한 sigmoid function이고, ρ_{id}는 매번 임의로 설정되는 0.0에서 1.0 사이의 값이다. 즉 위의 식에 따르면 v_{id}가 0.0에 가까울수록 x_{id}는 0의 값을 가질 확률이 높으며, 1.0에 가까울수록 1의 값을 가질 확률이 높게 된다.

4. 사전 중요도를 이용한 성능 향상

4.1 mRMR

mRMR[19] 방법은 상호 정보량(Mutual Information) 측정을 기반으로 하여 더 발전된 방법으로 제안된 것으로, 특징의 클래스와의 상호정보량을 관련성 (Relevance), 특징들 간의 상호정보량을 잉여성 (Redundancy) 으로 정의하여 두 척도를 조합하여 개별 특징의 우수성의 척도로 삼는 방법이다.

특징의 관련성은 다음과 같이 정의된다. 주어진 데이터가 하나의 샘플당 $X_1, X_2, ... X_m$의 특징과 클래스 C로 이루어지며, 특징 X_m은 $X_m = \{x_1, x_2, ... x_n\}$과 같은 특징값 분포를 가지고 클래스 C는 $C = \{c_1, c_2, ... c_k\}$와 같은 클래스 분포를 가진다고 할 때,

$$Relevance(X_m) = I(X_m; C)$$

와 같이 특징 X_m의 관련성 척도를 정의한다. 즉 특징 X_m과 클래스 C간의 상호 정보량이 특징의 관련성이 되는 것이다.

또한 특징의 잉여성은 각 특징들 간의 상호 의존도를 의미하는 것으로, 아래와 같이 정의한다.

$$Redundancy(X_m) = \frac{1}{m-1} \sum_{X_i \neq X_m} I(X_i; X_m)$$

즉 다른 특징들과 상호 정보량이 많은 특징일수록 특징들이 서로 그 특징값의 클래스에 대한 확률분포가 매우 유사하여 그 특징들이 함께 사용되어도 각 특징 하나씩만 사용하는 것과 비교하여 분류에 큰 도움이 되지 않는다고 보는 것이다. 다시 말해 잉여성이 적은 특징일수록 특징의 우수도가 높다고 볼 수 있다.

최종적으로 위의 관련성과 잉여성 척도를 결합하여 특징의 우수성을 판별하게 되는데, 관련성은 클수록 (Maximal - Relevance), 그리고 잉여성은 작을수록 (Minimal - Redundancy) 특징의 우수성은 크다고 볼 수 있다. 두 값을 다음과 같이 혼합하여 특징의 우수성을 최종적으로 도출하게 된다.

$$mRMR(X_m) = Relevance(X_m) - Redundancy(X_m)$$

mRMR 방법은 이렇게 특징의 관련성과 잉여성을 상호정보량 계산법을 통해 구하여 그 값들을 혼합해 특징의 우수성 척도로 사용하게 된다.

4.2 mRMR을 이용한 특징의 사전 중요도 설정

먼저 데이터 $D = \{x_1, x_2, ..., x_n\}$안의 모든 특징 x_i에 대해 각 특징의 mRMR 값을 구해 그 값을 m_i라 하자. 이를 토대로 전체 특징의 mRMR 값의 평균 $\overline{m}$와 표준편차 s를 구하여 데이터 D를 다음과 같은 세 부분집합으로 나눈다.

$$D_1 = \left\{ x_i \mid m_i - \overline{m} > \frac{s}{2} \right\}$$

$$D_2 = \left\{ x_i \mid -\frac{s}{2} \leq m_i - \overline{m} \leq \frac{s}{2} \right\}$$

$$D_3 = \left\{ x_i \mid m_i - \overline{m} < -\frac{s}{2} \right\}$$

이것은 즉 데이터 안의 특징들을 mRMR 값의 평균값을 중심으로 하여 평균적인 mRMR 값을 보이는 특징들, 평균값보다 월등히 큰 mRMR 값을 보이는 특징들, 그리고 평균값보다 월등히 작은 mRMR 값을 보이는 특징들로 나누게 되는 것이다. 이 구분을 이용해 GA와 PSO를 사용한 방법에서 각각 보조 지침으로 삼아 방법을 개선한다.

4.3 GA + mRMR

GA를 사용한 특징 선택에서는 비트 문자열들의 전체적인 교체가 일어나는 교배 단계에서는 개별적인 특징의 포함/미포함의 여부가 고려되기 힘들다. 개개의 특징, 즉 비트 문자열로 나타난 염색체의 각각의 비트에 대한 고려가 요구될 수 있는 단계는 교배 단계를 거친 뒤의 돌연변이 단계이다. 따라서 수정된 방법에서는 나머지 단계는 기존 GA와 마찬가지로 진행하도록 하고, 돌연변이 단계에서의 연산을 위의 보조 지침을 이용해 다음과 같이 수정한다.

```
Algorithm 3.1
Mutation in Feature Selection using GA + mRMR
N : Number of Population
 n : Number of Bits in Gene (=Number of Features)
G_i = {x_i1, x_i2, ..., x_in} : List of Features
P_m : Probability of Mutation (0 < P_m < 1)
...
/* Do Mutation */
For i = 1 to N
 For j = 1 to n
  get random value rand (0 < rand < 1)
   If (rand < P_m)
    If (x_ij ≡ 0)
     If (x_j ∈ D_3) then x_ij = 0      // No change
                    Else x_ij = 1
     Else If (x_ij ≡ 1)
     If (x_j ∈ D_1) then x_ij = 1      // No change
                    Else x_ij = 0
     Endif
 Endfor
Endfor
```

변형한 알고리즘을 간략하게 설명하면 다음과 같다. 돌연변이 단계에서 비트의 값을 바꿔야 하는 경우, 즉 해당 특징을 포함할지 안할지 결정하는 단계에서 mRMR 값에 의한 사전 중요도를 고려하는데, 비트가 0에서 1로 바뀌는 경우, 즉 해당 특징을 포함하지 않는 상태에서 포함하는 상태로 바꿔야 하는 돌연변이의 경우에 해당 특징이 D_3에 속하는 특징이라면 (mRMR 값이 작은 특징), 돌연변이를 행하지 않고 해당 특징이 포함되지 않는 상태를 그대로 유지하게 되는 것이다. 마찬가지로 반대의 경우, 즉 해당 특징을 포함하는 상태에서 포함하지 않는 상태로 바꿔야 하는 돌연변이의 경우에는 해당 특징이 D_1에 속하는 특징이라면 (mRMR 값이 큰 특징) 돌연변이를 행하지 않고 해당 특징이 포함된 특징 부분집합을 그대로 유지하게 된다. 그리고 그 외의 경우에는 일반적인 돌연변이 과정을 그대로 따르도록 한다.

이렇게 돌연변이 과정을 바꿈으로서 무작위로 행해지는 돌연변이 연산에 일정한 지침을 부여할 수 있게 된다. 기존의 돌연변이 연산을 적용하였을 때에는 해당 특징 부분집합에서 매우 높은 중요도를 보이는 특징이 돌연변이 연산에 의해 제거되어 버리거나, 중요도가 낮은, 혹은 분류 성능을 오히려 떨어뜨릴 수 있는 특징이 포함되어 버리는 현상을 막을 수 없었다. 그렇지만 이렇게 특징의 사전 중요도에 의한 지침으로 인해 돌연변이 단계에서 중요한 특징이 제거되거나 불필요한 특징이 포함되게 되는 것을 막을 수 있게 된다. 이를 통해 GA 연산에서 전체적으로 최적 해, 즉 가장 성능이 좋은 특징 집합을 찾아가는 데 걸리는 시간을 단축할 수 있고, 나아가 더 나은 성능의 특징 집합을 찾아내는 효과도 기대할 수 있다.

4.4 PSO + mRMR

GA와는 달리, PSO를 이용하는 방법은 각 파티클의 위치 벡터 요소 하나하나, 즉 특징 부분 집합의 각 개개의 특징에 대해 매번 고려하게 된다. 따라서 매번 각 특징에 대해 고려할 때마다 특징의 사전 중요도를 함께 고려하여 연산을 수행할 수 있다. 앞서 3.3 에서 설명한 이진 PSO 연산의 식을 마찬가지로 사전 지표를 이용해 수정하여 알고리즘화 하면 다음과 같다. *pbest*와 *gbest*를 찾는 등의 과정은 생략하고 매번 각 파티클의 위치 벡터 비트가 변하는 과정만을 고려하였다.

즉 최종적으로 비트의 값이 설정되는 과정에서 해당 비트가 1로 설정되는 상황인 경우 해당 특징이 D_3에 속하는 특징이라면 (mRMR 값이 작은 특징) 해당 비트를 0으로 설정하는 것으로 연산을 바꾸고, 반대로 해당 비트가 0으로 설정될 상황에서 해당 특징이 D_1에 속하는 특징이라면 (mRMR 값이 큰 특징) 해당 비트를 1로 설정하는 것으로 연산을 바꾸게 된다. 그리고 그 외의 경우에는 원래의 파티클 집단 최적화 연산과정을 그대로 따르게 한다. 이렇게 사전 중요도를 사용하여 중요도가 매우 높거나 낮은 특징은 확실히 포함 또는 불포함 하도록 만들어서, 분류에 도움이 되는 특징을 불포함하게 하거나, 또는 그 반대로 중요도가 매우 낮은 특징을 포함하게 하는 경우가 낮은 확률이지만 분명히 생기게 되는 현상을 막을 수 있다. 이러한 사전 중요도를 이용한 보완을 통해 전체적으로 최적 해, 즉 가장 성능이 좋은 특징 집합을 찾아가는 데 걸리는 시간을 단축할 수 있고, 나아가 더 나은 성능의 특징 집합을 찾아내는 효과도 기대할 수 있다.

Algorithm 3.2 Feature Selection using PSO + mRMR

T : Number of Iteration
N : Number of Particles
D : Number of Bits in Particle (=Number of Features)
$X_i^{(t)} = (x_{i1}, x_{i2}, ..., x_{iD})$: List of Features in ith particle at time t
c_1, c_2 : cognitive parameter

For $t = 1$ to T
 For $i = 1$ to N
  For $d = 1$ to D
   $v_{id}^{(t)} = v_{id}^{(t-1)} + c_1(p_{id} - x_{id}^{(t-1)}) + c_2(p_{gd} - x_{id}^{(t-1)})$
   $s(v_{id}) = \dfrac{1}{1+\exp(-v_{id})}$
   Get random value ρ_{id} $(0.0 \le \rho_{id} \le 1.0)$
    If ($\rho_{id} < s(v_{id}(t))$)
     If $(x_j \in D_3)$ then $x_{ij}(t) = 0$
     Else $x_{id}(t) = 1$
    Else If ($\rho_{id} > s(v_{id}(t))$)
     If $(x_j \in D_1)$ then $x_{ij}(t) = 1$
     Else $x_{id}(t) = 0$
  Endfor
 Endfor
Endfor

4. 실험 및 결과

4.1 실험 데이터 및 실험 방법

본 논문에서 실험을 위해 사용된 데이터는 총 20개로, 모든 데이터 셋은 UCI Machine Learning Repository 에 있는 실제 데이터들이다[20]. 특징 선택에 대한 성능 비교가 목적이기 때문에 모든 데이터는 특징의 수가 충분히 많은 것들로 선택하였고, 특징값의 형식도 이산형 데이터와 연속형 데이터들이 고루 분포되도록 실험 데이터를 구성하였다. <표 1>에서 실험에 사용한 데이터들에 대해 요약해 놓았다.

먼저 선 실험으로 위의 데이터들을 특징 선택을 하지 않은 상태에서 분류 성능을 측정해 보았다. 성능 측정을 위해 사용한 분류기는 베이즈 이론 (Bayes Theory)을 바탕으로 한 나이브 베이즈 분류기(Naive Bayes Classifier, NB)[21]와 지지 벡터 머신 (Support Vector Machine)을 구현한 선형 커널(Linear Kernal)을 사용한 SMO 분류기[22], 그리고 결정 트리 이론 (Decision Tree Theory)을 바탕으로 한

〈표 1〉 실험에 사용한 데이터

데이터	특징 수	클래스 수	샘플 수
Audiology	69	24	226
Dermatology	33	6	366
Musk	166	2	475
Spambase	57	2	4601
Arrhythmia	277	16	452
Ionosphere	34	2	351
Waveform	21	3	5000
Sonar	60	2	208
Image Segmentation	19	7	2310
Flag	28	6	194
Hepatitis	19	2	155
Lung Cancer	56	3	32
Promoter	56	2	106
Splice	60	3	3190
Optdigits	64	10	3823
SpectF	44	2	267
Connect-4	42	3	691
Water Treatment	38	13	527
Isolet	617	26	6236
HDR Multifeature	649	10	2000

〈표 2〉 원본 데이터의 분류 정확도

데이터	분류정확도(%)		
	NB	C4.5	SMO
Audiology	73.45	77.88	81.86
Dermatology	97.54	90.98	95.36
Musk	73.68	78.95	82.74
Spambase	79.29	92.61	90.42
Arrhythmia	59.07	66.37	67.70
Ionosphere	84.62	88.60	89.17
Waveform	78.20	75.60	81.22
Sonar	65.38	69.23	73.56
Image Segmentation	79.87	87.84	85.80
Flag	56.19	70.62	64.43
Hepatitis	85.16	79.35	85.81
Lung Cancer	62.50	59.38	50.00
Promoter	82.08	82.08	82.08
Splice	91.25	92.60	84.55
Optdigits	91.66	89.62	98.14
SpectF	68.54	73.78	79.78
Connect-4	54.85	63.68	63.82
Water Treatment	74.76	70.02	78.75
Isolet	84.35	83.24	96.81
HDR Multifeature	95.35	94.60	98.40

C4.5 분류기이다[23]. 이 3개의 분류기에 대해 각각 분류 성능을 측정하였다. 3개의 분류기 모두 공개 데이터마이닝 실험 소프트웨어인 WEKA3.5[24]에 구현된 것을 사용하였다. 또한 분류 성능 측정 방식은 10겹 교차검증 방식 (10-Fold Cross-Validation)을 사용하였다.

이를 토대로 본 실험에서는 먼저 GA와 PSO를 사용한 특징 선택 방법이 전체 데이터를 사용했을 때보다 더 좋은 분류 성능을 보여주는 것을 확인하도록 하였다. 그리고 논문에서 최종적으로 제시하는 방법인 mRMR과 생태계 모방 알고리즘을 결합한 방법의 성능을 실험을 통해 확인한다. 이 실험에선 기존의 GA와 PSO를 사용한 특징 선택 방법과 비교하여 mRMR을 이용한 특징의 사전 중요도를 보조 지침으로 한 보완된 GA와 PSO 방법이 얼마나 나은 성능을 보여주는지를 확인한다.

GA와 PSO 모두 각각 40개의 염색체/파티클을 이용하여 20회의 반복(진화)를 거쳐 최종적으로 가장 적합도가 좋은 특징 부분집합을 취한다. 적합도 함수로 분류기의 분류성능을 직접 사용하도록 설정했기 때문에 따로 다시 분류성능을 측정할 필요는 없다.

기존의 GA와 PSO를 사용한 방법들의 분류 성능과 사전 중요도를 사용한 GA와 PSO를 이용한 분류 성능의 비교는 <표 3-5>와 같다. 각각 기존 GA와 mRMR과 결합된 GA

〈표 3〉 기존 GA, PSO와 사전 중요도를 이용한 방법과의 성능 비교 (NB)

데이터	GA	GA + mRMR	PSO	PSO + mRMR
Audiology	76.55	77.43	78.32	77.88
Dermatology	99.18	99.18	99.18	99.45
Musk	79.58	83.37	81.68	84.00
Spambase	89.52	89.11	90.11	91.31
Arrhythmia	67.70	68.14	69.69	72.35
Ionosphere	90.88	91.45	91.74	92.02
Waveform	79.36	79.14	79.48	79.48
Sonar	75.48	76.44	76.92	79.81
Image Segmentation	84.85	84.85	84.85	84.85
Flag	70.62	70.10	72.16	73.20
Hepatitis	89.68	89.68	90.32	90.32
Lung Cancer	84.38	84.38	87.50	90.63
Promoter	92.45	93.40	94.34	95.28
Splice	91.38	91.44	91.63	91.97
Optdigits	92.73	93.60	93.23	93.70
SpectF	79.40	79.40	79.03	79.40
Connect-4	66.71	66.70	68.74	69.03
Water Treatment	78.94	79.13	79.32	77.99
Isolet	89.10	89.57	89.64	89.64
HDR Multifeature	96.35	96.25	96.70	97.30

〈표 4〉 기존 GA, PSO와 사전 중요도를 이용한 방법과의 성능 비교 (C4.5)

데이터	GA	GA + mRMR	PSO	PSO + mRMR
Audiology	78.32	78.32	78.32	78.32
Dermatology	95.90	95.90	95.90	95.90
Musk	84.00	85.26	88.00	89.26
Spambase	93.13	93.70	93.81	93.61
Arrhythmia	72.35	72.35	73.23	73.23
Ionosphere	92.02	92.02	91.17	92.31
Waveform	76.84	76.84	77.64	77.64
Sonar	78.85	78.85	84.13	84.13
Image Segmentation	88.10	88.14	88.23	88.23
Flag	73.71	74.23	74.23	75.26
Hepatitis	85.81	84.52	85.81	85.81
Lung Cancer	71.88	71.88	71.88	71.88
Promoter	86.79	86.79	86.79	86.79
Splice	94.48	93.98	94.70	94.61
Optdigits	90.03	89.69	90.40	91.43
SpectF	83.52	83.52	86.14	84.64
Connect-4	70.19	70.62	70.33	70.62
Water Treatment	72.87	73.06	76.28	74.95
Isolet	83.55	85.55	84.56	86.57
HDR Multifeature	96.40	97.10	96.65	96.65

〈표 5〉 기존 GA, PSO와 사전 중요도를 이용한 방법과의 성능 비교 (SMO with Linear Kernel)

데이터	GA	GA + mRMR	PSO	PSO + mRMR
Audiology	85.84	85.84	85.84	87.15
Dermatology	98.63	98.91	98.91	98.91
Musk	85.26	85.05	85.89	85.89
Spambase	90.09	90.39	90.39	90.48
Arrhythmia	73.01	75.45	74.25	75.45
Ionosphere	90.60	90.31	92.31	91.45
Waveform	81.42	81.42	81.60	81.60
Sonar	81.73	82.21	84.62	83.65
Image Segmentation	86.10	86.10	86.15	86.15
Flag	66.49	69.55	69.07	71.65
Hepatitis	88.39	86.45	89.68	89.68
Lung Cancer	78.13	84.38	87.50	87.50
Promoter	96.23	93.40	95.28	95.28
Splice	85.36	85.36	85.17	85.30
Optdigits	98.17	98.22	98.33	98.38
SpectF	79.40	79.40	79.40	79.40
Connect-4	64.83	64.83	64.83	64.83
Water Treatment	79.51	80.08	79.89	81.21
Isolet	96.75	96.90	96.75	97.43
HDR Multifeature	99.10	99.45	99.53	99.53

(GA + mRMR), PSO와 mRMR과 결합된 PSO(PSO + mRMR)를 서로 비교하였으며, 성능이 더 높은 쪽에 밑줄로 표시하였다.

보이는 바와 같이, GA와 PSO를 사용하여 특징 선택을 했을 경우 전체 데이터를 사용했을 때보다 분류 성능 면에서 더 좋아지는 것을 확인할 수 있고, 또한 대부분의 경우 mRMR을 이용해 사전 중요도를 설정하여 개선한 방법은 GA와 PSO 양쪽에서 기존의 GA와 PSO방법보다도 더 나은 분류 정확도를 나타내었다. 전체적으로 보았을 때는 PSO + mRMR 쪽이 가장 나은 분류 정확도를 보여, 분류 성능 면에서 가장 나은 능력을 보인다고 할 수 있다.

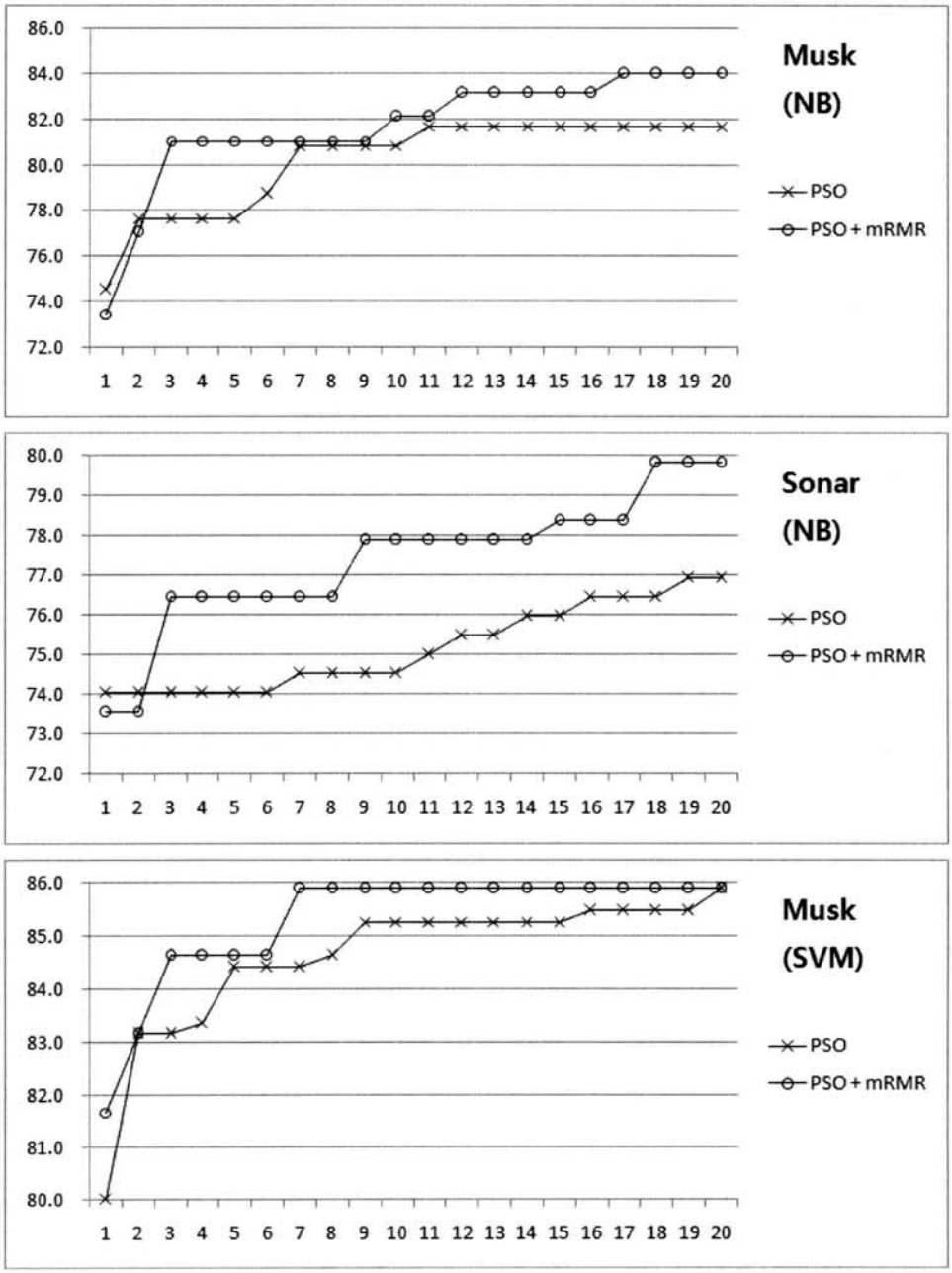
두 번째로, 세대를 반복해 가면서 분류정확도가 상승하는 추세를 각 방법별로 비교해 보도록 한다. 이 비교는 어느 방법이 더 빠른 시간 안에 최적의 특징 부분집합을 찾아낼 수 있는지의 여부를 판단하기 위한 비교로, 전체 데이터 중 몇 몇 데이터에 대하여 매 세대마다 도출되는 최적 특징 부분집합의 분류 성능을 그래프로 표현하여 비교해 보았다. 가로축은 세대의 반복 회수를, 세로축은 적합도 함수, 즉 분류 성능(%)을 나타낸다.

그래프를 보면 알 수 있듯이, 20회의 정해진 반복을 거치는 과정에서 mRMR을 사용한 방법이 기존 방법보다 더 빠르게 최고 분류 성능에 도달한다. 즉 mRMR을 사용한 방법이 빠르게 분류 성능을 증가시켜 주므로, 기존 방법에 비해 더 적은 시간 (반복) 안에 최적의 특징 조합을 찾아내는 데 유용하다는 것을 알 수 있다.

5. 결 론

지금까지 특징 선택의 전반적인 이론과 함께 생태계 모방 알고리즘을 이용하여 특징 선택 문제를 해결하는 방법 중 유전자 알고리즘(GA)과 파티클 집단 최적화(PSO)를 이용한 래퍼 방식의 특징 선택 방법에 대해 생각해 보았다. 그리고 최종적으로 기존의 특징 선택 방법 중 하나인 상호 정보량 기반의 mRMR 방법을 결합하여 특징의 개별 중요도를 사전 지표로 사용하여 위의 두 생태계 모방 알고리즘을 이용한 방식을 더 발전시킨 특징 선택 방법에 대해 제안하고, 실제 데이터를 이용해 제안한 방식의 성능을 검증해 보았다.

실험을 통해 확인할 수 있었듯이, GA와 PSO의 두 생태계 모방 알고리즘을 이용한 래퍼 방식의 특징 선택 방법은 데이터의 분류 성능 향상에 많은 도움을 주었다. 분류 성능



(그림 2) 분류성능(적합도)의 변화 추세 비교

을 잘 발휘할 수 있는 특징 부분집합을 잘 생성해 주어 전체 데이터를 사용했을 때보다 좋은 분류 성능을 보여주는 것을 확인할 수 있었다. 이는 최적화된 해를 찾는 문제에서 뛰어난 성능을 보이는 생태계 모방 알고리즘의 장점이 잘 발휘된 결과라고 할 수 있다.

또한 거기에 더해 mRMR 방법을 이용하여 개별적인 특징의 우수성을 판별하여 생태계 모방 알고리즘 내에서의 연산의 지침으로 삼아 기존의 생태계 모방 알고리즘을 이용한 방식을 더 효율적으로 개선할 수 있는 방법을 생각해 보았고, 이 또한 실험을 통해서 기존의 생태계 모방 알고리즘만 사용한 방식보다 그 성능 면에서 대체로 우수함을 확인할 수 있었다.

물론 각 방법들의 사전 변수 설정이나 분류 알고리즘의 특성 등을 고려하여야 하지만, 우리가 일반적으로 데이터를 분류하는 문제를 접할 때 GA또는 PSO와 mRMR을 결합한 특징 선택 방법을 사용하면 분류 알고리즘에 상관없이 충분히 좋은 성능을 보여줄 것이라고 기대해도 좋을 것이다.

차후에는 사전 설정 변수의 값 등 여러 조건들을 더 다양하게 고려한 세밀한 실험을 통한 심화된 성능 비교 분석이 요구된다. 또한 GA와 PSO외에도 다른 생태계 모방 알고리즘을 사용하여 성능을 지금의 결과와 비교하여 보는 것도 의미있는 향후 과제라고 할 수 있겠다.

참 고 문 헌

[1] Blum, A. and Langley, P., "Selection of Relevant Features and Examples in Machine Learning," Artificial Intelligence, Vol.97, No.1-2, pp.245-271, 1997.

[2] Liu, H. and Yu, L., "Toward Integrating Feature Selection Algorithms for Classification and Clustering," IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.4, pp.491-502, 2005.

[3] Jain, A. and Zongker, D., "Feature Selection : Evaluation, Application, and Small Sample Performance," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.2, pp.153-158, 1997.

[4] Zhao, J., Wang, G., Wu, Z., Tang, H. and Li, H., "The Study on Technologies for Feature Selection," Proceedings of the First International Conference on Machine Learning and Cybernetics, pp.689-693, 2002.

[5] Mitchell, M., An Introduction to Genetic Algorithms, MIT PRESS, 1996.

[6] Bautista, M. and Vila, M., "A Survey of Genetic Feature Selection in Mining Issues," Proceedings of the Congress on Evolutionary Computation, Vol.2, pp.1314-1321, 1999.

[7] Yang, J. and Honavar, V., "Feature Subset Selection Using a Genetic Algorithm," IEEE Intelligent Systems, Vol.13, pp.44-49, 1998.

[8] Liu, Y., Qin, Z., Xu, Z. and He, X., "Feature Selection with Particle Swarms," Proceedings of International Symposium on Computational and Information Science, pp.425-430, 2004.

[9] Firpi, A. and Goodman, E., "Swarmed Feature Selection," Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop, pp.112-118, 2004.

[10] Yan, Z. and Yuan, C., "Ant Colony Optimization for Feature Selection in Face Recognition," Lecture Note in Computer Science, Vol.3072, SPRINGER, 2004.

[11] Galbally, J., Fierrez, J., Freire, M.R. and Ortega-Garcia, J., "Feature Selection Based on Genetic Algorithms for On-Line Signature Verification," Proceedings of IEEE Workshop on Automatic Identification Advanced Technologies, pp.198-203, 2007.

[12] Bello, R., Gomez, Y., Nowe, A. and Garcia, Maria M. "Two-Step Particle Swarm Optimization to Solve the Feature Selection Problem," Proceedings of Seventh International Conference on Intelligent Systems Design and Applications, pp.691-696, 2007.

[13] Geetha, K., Thanushkodi, K. and Kumar, A.K. "New Particle Swarm Optimization for Feature Selection and Classification of Microcalcifications in Mammograms," Proceedings of International Conference on Signal Processing, Communications and Networking, pp.458-463, 2008.

[14] Muni, D.P., Pal, N.R. and Das, J. "Genetic programming for simultaneous feature selection and classifier design," IEEE Transactions on Systems, Man, and Cybernetics, Vol.36, pp. 106-117, 2006.

[15] Goldberg, D., Genetic Algorithms in Search, Optimization, and Machine Learning, ADDISON-WESLEY, 1989.

[16] Kennedy, J. and Eberhart, R., Swarm Intelligence, MORGAN KAUFMANN, 2001.

[17] Kennedy, J. and Eberhart, R., "Particle Swarm Optimization," Proceedings of the Conference on Neural Networks, pp.1942-1948, 1995.

[18] Engelbrecht, A., Fundamentals of Computational Swarm Intelligence. WILEY, 2005.

[19] Peng, H., Long, F. and Ding, C., "Feature Selection based on Mutual Information : Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.8, pp.1226-1238, 2005.

[20] Blake, C. and Merz, C., UCI Repository of Machine Learning Database, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[21] Mitchell, T., Machine Learning, MCGRAW-HILL, 1997.

[22] Platt, C., "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in Advances in Kernel Methods: Support Vector Learning, pp.185-208, 1999.

[23] Quinlan, J., C4.5: Programs for Machine Learning, MORGAN KAUFMANN, 1993.

[24] WEKA3.5, http://www.cs.waikato.ac.nz/~ml/

윤 철 민

e-mail : toro83@nate.com

2006년 서강대학교 컴퓨터학과(학사)

2008년 서강대학교 대학원 컴퓨터학과 (공학석사)

2008년~현 재 (주)다이퀘스트 기술연구소 선임연구원

관심분야 : 기계학습, 데이터마이닝, 진화 연산 등

양 지 훈

e-mail : yangjh@sogang.ac.kr

1987년 서강대학교 전자계산학과(학사)

1989년 아이오와 주립대학교 대학원 컴퓨터학과(공학석사)

1999년 아이오와 주립대학교 대학원 컴퓨터학과(공학박사)

1999년~2000년 HRL Lab. LLC., Research Staff Member

2000년~2002년 SRA International, Inc., Professional Staff Member

2002년~현 재 서강대학교 컴퓨터학과 부교수

관심분야 : 기계학습, 데이터마이닝, 인공지능, 생물정보학 등