생태계 모방 알고리즘을 이용한 특징 선택 방법들의 성능 비교 분석에 대한 연구

윤철민, 양지훈
서강대학교 컴퓨터공학과
e-mail : cmyun@sogang.ac.kr

An Experimental Comparison of Feature Subset Selection Methods using Bio-Inspired Algorithms

Chulmin Yun, Jihoon Yang
Dept. of Computer Science, Sogang University

요 약

패턴 인식 문제를 푸는데 있어 특징 선택을 해주는 것은 패턴 인식의 성능 향상을 위해 중요한 과정 중 하나이다. 본 연구에서는 대표적인 생태계 모방 알고리즘 2 가지를 선택하여 특징 선택 문제에 적용하여 보고, 그 성능을 비교 분석하였다. 데이터의 특징을 줄여주는 기능과 패턴 인식 성능의 향상 여부를 중심으로 평가하였으며, 이를 통해 생태계 모방 알고리즘이 특징 선택 문제에 효과적으로 사용될 수 있는지에 대해 논의해보고, 두 방법의 장단점과 특징에 대해 생각해 본다.

1. 서론

특징 선택 (Feature Subset Selection) 은 패턴 인식 문제에서 중요한 이슈 중 하나로, 복잡하고 방대한 데이터에서 패턴 인식에 필요한 주요 속성들을 골라내어 데이터의 크기를 줄여주게 되며, 이로 인해 패턴 인식기의 수행 시간과 성능을 향상시켜 준다. 이러한 특징 선택 방법은 오랜 시간에 걸쳐 다양한 방법들이 개발되어 왔으며, 현재도 꾸준히 새로운 방법들이 제시되고 있다.

생태계 모방 알고리즘은 생태계에서 일어나는 생물체들의 행동 습성 등을 관찰하여 개발된 알고리즘으로, 유전자의 변이, 진화에서 아이디어를 빌린 유전자 알고리즘, 개미의 습성을 알고리즘화 한 개미 군집 최적화 알고리즘, 새떼 등의 이동 모습에서 착안한 미립자 집단 최적화 알고리즘 등이 있다.

본 연구에서는 이러한 생태계 모방 알고리즘을 특징 선택 문제에 적용하여 그 성능에 대해 비교 분석해 보기로 한다. 유전자 알고리즘과 미립자 집단 최적화 알고리즘을 이용하여 주어진 데이터에 대해 특징 선택을 수행한 뒤 그에 따른 분류 성능 변화에 대해 관찰하고, 두 방법간의 성능 차이에 대해 살펴볼 것이다.

2. 생태계 모방 알고리즘을 이용한 특징 선택

실험을 위해 사용한 생태계 모방 알고리즘은 유전자 알고리즘과 미립자 집단 최적화 알고리즘의 두 가지이며, 각각의 방법에 대한 간략한 소개와 방법들을 특징 선택에 적용하는 방법은 다음과 같다.

(1) 유전자 알고리즘 (GA)

유전자 알고리즘 (Genetic Algorithm, GA) 은 생물체의 유전자가 변이하고 진화하는 과정을 알고리즘화 한 것으로, 풀고자 하는 문제에 대한 가능한 해들을 염색체로 표현한 다음, 이들을 점차적으로 변형함으로써 더 좋은 해들을 생성하게 된다. 이 과정에서 선택(Selection), 교배(Crossover), 돌연변이(Mutation) 등의 연산을 사용하여 유전자를 변화시키게 되는데, 매 세대 마다 유전체가 어느 정도나 좋은 해인지 평가하는 적합도 함수 (Fitness Function)를 적용하여 가장 적합도가 높은 유전체를 다음 세대에 남겨두어 사용하게 한다.

특징 선택 문제에 GA 를 적용할 때는 하나의 염색체가 하나의 가능한 특징들의 조합이 된다. 염색체는 전체 특징 수만큼의 길이를 가지는 이진 비트 문자열로 구성되며 문자열의 각 자리에 해당하는 특징이 포함되었을 때는 1, 포함되지 않았을 때는 0 의 값을 가진다.

(2) 미립자 집단 최적화 (PSO)

미립자 집단 최적화 (Particle Swarm Optimization, PSO) 는 새, 벌 등의 군집 생활을 하는 생물체들의 이동 모습과 그 원리를 알고리즘화 한 것이다. 집단은 여러 개의 미립자 (Particle) 로 이루어져 있으며, 각각의 미립자들의 위치 (Position) 는 풀고자 하는 문제에 대한 가능한 각각의 해가 된다. 각각의 미립자들은 매번 반복하여 자신의 위치, 즉 해를 변화시키게 되는데, 이 과정에서 모든 미립자들의 위치 중에 가장 좋았던, 즉 적합도가 높았던 위치 (Global Best Position) 와 각 미립자들의 위치변화 중 가장 적합도가 높았던 위치

(Local Best Position) 를 고려하여 위치를 변화시킨다. 이 과정을 반복하게 되면서 점점 적합도가 높아지는 위치로 미립자들은 이동하게 된다. 즉 해가 점점 좋은 성능을 발휘하는 쪽으로 변화되게 된다.
특징 선택 문제에 PSO 를 적용할 때에는 GA 와 비슷한 방법으로 각 미립자의 위치는 각각 하나의 가능한 특징 조합이 된다. 즉 미립자의 위치가 변화하는 것은 특징의 조합이 변화하게 되는 것을 의미하게 된다.

3. 실험

(1) 데이터
실험을 위해 총 5 개의 데이터를 사용하였다. 모든 데이터는 UCI Machine Learning Repository 에 있는 실제 데이터들이다. 각 데이터들에 대한 간략한 설명은 <표 1> 에 요약되어있다.

<표 1> 실험에 사용한 데이터에 대한 요약

	특징 수	샘플 수	클래스 수
Arrhythmia	277	452	16
Image Segmentation	19	2310	7
Waveform	21	5000	3
Sonar	60	208	2
Ionosphere	34	351	2

(2) 적합도 함수 설정
GA 와 PSO 모두 적합도 함수 (Fitness Function)을 설정하여 해의 적합도를 판별하게 된다. 이 실험에서는 각각 염색체와 미립자가 가지고 있는 특징 조합대로 재구성된 데이터를 분류 알고리즘에 적용하여 그 분류 정확도를 해당 염색체나 미립자의 적합도로 사용하였다. 분류 알고리즘으로는 베이즈 결정 이론을 기반으로 한 Naïve Bayes 방법이 사용되었고, 10 겹 교차검증 (10-fold Cross-Validation) 을 이용하여 분류 정확도를 측정하였다.

4. 결과 분석

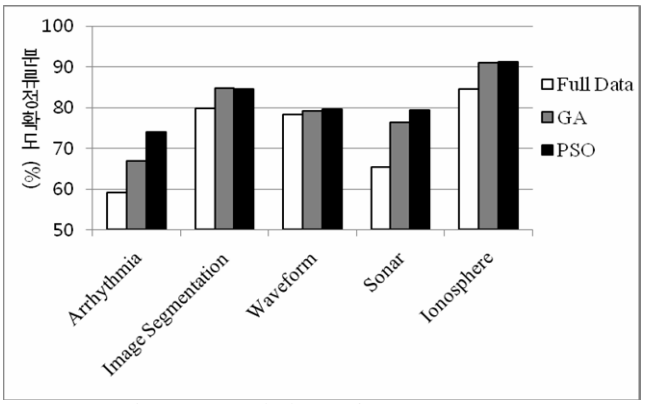
먼저, 특징 선택을 하지 않은 상태, 즉 모든 특징이 포함된 상태에서 각 데이터의 정확도를 GA 와 PSO 의 적합도 측정과 같은 방법으로 (Naïve Bayes, 10-fold Cross-Validation) 측정하였다. 그리고 GA 와 PSO 두 방법을 통해 나온 가장 적합도가 높은 염색체 혹은 미립자의 특징 수와 그 때의 적합도를 구해 비교하였다.

(1) 선택된 특징의 수
두 방법 모두 실험에 사용된 모든 데이터에 대해 전체 데이터의 특징 수 보다 더 적은 수의 특징을 선택해 주는 것을 <표 2> 를 통해 확인할 수 있다.
두 방법이 모두 데이터의 크기를 줄여 주는 역할은 문제없이 잘 수행해 주고 있다는 것을 확인할 수 있는 결과이다. 특히 PSO 방법에 비해 GA 방법이 약간 더 적은 수의 특징을 선택해 주는 것을 알 수 있다.

<표 2> GA 와 PSO 를 통해 선택된 특징 수 비교

특징 수 / 데이터	전체	GA	PSO
Arrhythmia	277	103	118
Image Segmentation	19	9	12
Waveform	21	16	16
Sonar	60	21	22
Ionosphere	34	15	18

(2) 분류 정확도의 비교
(그림 1) 을 통해 모든 특징이 선택되었을 때의 데이터와 GA 와 PSO 를 통해 나온 가장 적합도가 높은 특징 조합을 가지는 데이터의 Naïve-bayes 분류 정확도를 비교해 볼 수 있다.



(그림 1) 분류정확도 비교 (Naïve-bayes)

GA 와 PSO 의 두 특징 선택 방법 다 모든 데이터에대해 모든 특징을 포함하였을 때보다 더 좋은 분류 정확도를 보여 주는 것을 볼 수 있다. 이로서 두 방법 모두 특징 선택에 적용하였을 때 충분히 효과를 발휘할 수 있는 방법이라는 것을 알 수 있다.
두 방법이 많은 성능 차이를 보이고 있지는 않지만, 몇몇의 데이터, 특히 Arrhythmia 와 Sonar 데이터에 대해서는 PSO 방법이 GA 보다 좋은 성능을 보이는 것을 알 수 있다. 이는 위의 선택된 특징 수 비교 결과와 함께 생각하였을 때, PSO 가 GA 보다 조금 더 많은 수의 특징을 선택하였지만 분류 성능은 더 나은 것을 보아 PSO 가 좀더 분류에 밀접하게 관여하는 특징들을 더 잘 골라내 준다고 볼 수 있다.

5. 결론

지금까지 대표적인 생태계 모방 알고리즘인 유전자 알고리즘 (GA) 과 미립자 집단 최적화 알고리즘 (PSO) 을 특징 선택에 적용하여 그 성능을 선택되는 특징의 수와 분류 정확도 향상 측면에서 비교해 보았다. 두 방법 모두 특징 선택을 통해 특징의 수를 줄이고 정확도의 향상을 가져오는 것을 확인할 수 있었다. 또한 두 방법간의 성능 비교에서 PSO 가 GA 보다는 분류 성능에 있어 약간 더 우위를 점한다는 것을 보였다. 그러나 PSO 와 GA 모두 실행에 있어 사전에 설정되어야 하는 여러 변수들이 존재하고, 그 변수의 설정을 어떻게 하느냐에 따라서 결과는 다르게 나올 수 있음을 고려해야 할 것이다.
GA 와 PSO 는 생태계 모방 알고리즘 중 대표적으로

알려진 일부 알고리즘들이다. 차후에는 이 두 가지 방법 외에 다른 다양한 생태계 모방 알고리즘, 예를 들면 개미 군집 최적화 알고리즘 등의 방법을 더 추가하여 마찬가지로 특징 선택에 적용하여 그 성능을 기존의 두 방법과 비교하는 연구도 필요할 것으로 생각한다.

참고문헌

[1] Avrim L. Blum and Pat Langley “Selection of Relevant Features and Examples in Machine Learning” ACM Artificial Intelligence Archive, Vol. 97, Issue 1-2, 245-271, 1997.

[2] Isabelle Guyon and Andre Elisseeff “An Introduction to Variable and Feature Selection” Journal of Machine Learning Research 3 , 1157-1182, 2003.

[3] David E. Goldberg “Genetic Algorithms in Search, Optimization & Machine Learning”, Addison Wesley, 1989.

[4] James Kennedy and Russell C. Eberhart, “Swarm Intelligence”, Morgan Kaufmann, 2001.

[5] Yu Liu, Zheng Qin, Zenglin Xu and Xingshi He “Feature Selection with Particle Swarms” Lecture Note in Computer Science, Vol. 3314, 425-430, 2004.

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

[7] UCI ML Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html