DOI: 10.3745/KIPSTB.2010.17B.2.157

분산 유전 알고리즘에서 자동 마이그레이션 조절방법

이 현 정[†] · 나 용 찬[††] · 양 지 훈[†††]

요 약

본 논문에서는 분산된 거대한 네트워크상의 데이터에서 유용한 정보를 추출하는 새로운 마이그레이션 조절방법을 이용한 유전 알고리즘을 제안한다. 제안된 알고리즘의 주된 아이디어는 부분 개체군 사이에서 개체들의 이동에 필요한 파라미터들을 적응적으로 결정하는 것이다. 또 이동된 개체들이 새로운 부분 개체군에서 도태되지 않고 적응 할 수 있기 위한 방법을 제시한다. UCI 기계학습 관련 데이터 셋에서 중앙 집중적 단일 유전 알고리즘과 제안된 알고리즘을 비교하기 위해 여섯 개의 데이터를 사용했다. 결론적으로 분산 유전 알고리즘을 적용한 특징 부분집합이 단일 유전 알고리즘을 적용한 것 보다 좋은 성능을 보였다.

키워드 : 유전 알고리즘, 분산 알고리즘, 자동 마이그레이션

Distributed Genetic Algorithm using Automatic Migration Control

Hyunjung Lee[†] · Yongchan Na[††] · Jihoon Yang[†††]

ABSTRACT

We present a new distributed genetic algorithm that can be used to extract useful information from distributed, large data over the network. The main idea of the proposed algorithms is to determine how many and which individuals move between subpopulations at each site adaptively. In addition, we present a method to help individuals from other subpopulations not be weeded out but adapt to the new subpopulation. We used six data sets from UCI Machine Learning Repository to compare the performance of our approach with that of the single, centralized genetic algorithm. As a result, the proposed algorithm produced better performance than the single genetic algorithm in terms of the classification accuracy with the feature subsets.

Keywords : Genetic Algorithm, Distributed Algorithm, Automatic Migration

1. 서 론

오늘날 데이터는 네트워크화 된 세계에서 다른 장소들에 만들어져 쌓이고 유지된다. 만약 우리가 그러한 데이터를 마이닝(Mining) 하기 원한다면 중앙으로 모아서 (Centralized) 계산하는 방법과 분산된(Distributed) 형태로 계산하는 방법 중 선택해야한다. 전자는 훈련 정확도(Training Accuracy)와 같은 확실한 기준 하에 전체 데이터에 대한 가장 좋은 마이닝 결과가 산출될 것을 기대할 수 있지만, 그것은 모든 데이터를 중앙으로 옮기는 것에 대한 오버헤드를 야기할 수 있으며 따라서 많은 실제 응용들에서 그 접근은 타당하지 않을 수 있다. 후자는 오직 지역 데이터(Local Data)만이 각각의 장소에서 고려되므로 전자에 비해 오버피팅(Overfitting)의 가능성이 적다. 그러므로 분산 데이터마이닝 또는 분석은 매우 의미 있다[1].

유전 알고리즘(Genetic Algorithm, GA)은 생물학에서 나온 전역 탐색 휴리스틱 알고리즘(Global Search Heuristic Algorithm) 중 하나로서 여러 가지 최적화 문제들에 적용되어왔다[2-4]. 분산유전 알고리즘(Distributed GA, DGA)은 시간, 공간 복잡도 오버헤드의 한계들을 완화하고 단일버전(Centralized GA, CGA)의 일반화 능력을 강화하기 위한 접근이다[1, 6-7].

DGA는 GA에 필요한 일반적인 파라미터들에 더하여 마이그레이션(Migration) 간격, 이주자들의 숫자, 이주자들의 선택과 교체 방법들, 그리고 환경의 네트워크 토폴로지와 같은 마이그레이션 정책을 위한 부가적 파라미터들을 필요로 한다[5]. 본 논문에서는 마이그레이션에 관련된 파라미터들의 셋(Set)을 정의하고 환경에 따라 자동적으로 그것들의 값을 조절하는 새로운 DGA를 제안한다.

본 논문의 나머지 장들은 다음과 같이 구성된다. 제 2장

※ 이 연구는 서강대학교 특별연구비 및 한국연구재단 일반연구자 지원사업의 지원에 의한 것임.
† 정 회 원 : 삼성전자 DMC 연구소 재직
†† 준 회 원 : 서강대학교 컴퓨터공학과 박사과정
††† 정 회 원 : 서강대학교 컴퓨터공학과 부교수
논문접수 : 2009년 9월 21일
수 정 일 : 1차 2009년 11월 20일, 2차 2009년 12월 28일
심사완료 : 2010년 2월 15일

은 GA와 DGA에 대한 배경지식 및 본 논문에서 그 응용으로서 이용된 특징 부분 선택에 대해 간략하게 설명한다. 제 3장에서는 제안된 자동 마이그레이션 조절을 이용한, 새로운 DGA 알고리즘과 그것의 응용인 특징부분 선택에 대한 세부적인 묘사를 한다. 제 4장에서는 우리가 제안한 방법의 성능을 평가하기 위해 여러 가지 실세계 데이터에서 실험한 결과를 보여주며, 마지막으로 5장은 요약과 향후 연구에 대한 토론으로 마친다.

2. 관련 연구

본 장에서는 일반적인 유전 알고리즘과 알려진 병렬 유전 알고리즘 그리고 특징선택(Feature Selection)에 관해 살펴보도록 한다.

2.1 유전 알고리즘

유전 알고리즘은 John Holland[2]에 의해 개발된 통계적 탐색 방법의 일종 이다. 이는 생태계 모방 알고리즘의 하나로서, 생명체의 유전과 진화의 과정에서 "적자생존"의 원리를 인공적으로 모사하여 최적의 해(Solution)를 찾는 최적화 문제(Optimization Problem)에 주로 이용된다. 유전 알고리즘은 선택(Selection), 교배(Crossover), 돌연변이(Mutation)의 단계로 구성되어 있다. 선택단계에서는 더 높은 적합도를 갖는 염색체(Chromosome)가 선택이 되고 선택된 두 개의 부모 염색체가 교배를 한다. 이때 교배 점은 확률적으로 결정한다. 돌연변이 단계에서는 교배에 의해 재조합 된 염색체의 일부를 확률에 의해 바꾸어준다. 이렇게 각 염색체는 적합도가 높을수록 많은 유전을 남기고 과정은 일정한 성능이 나오거나 일정횟수를 진행 한 후 최적 해를 만들어낸다.

2.2 유전 알고리즘의 병렬 모델들

현재까지 가장 잘 알려진 모델은 분산된 모델(Distributed Model, dGA)과 셀 방식의 모델(Cellular Model, cGA)이다[5-7]. 분산 모델은 굵은 입자 모델(Coarse-Grained model)이라고도 하며 전체 개체군을 더 작은 부분개체군들로 나누어 각각을 프로세서에 할당한다. 각각의 프로세서에서 부분개체군은 동시에, 독립적으로 진화하게 된다. 또한 프로세서 간의 통신이 허용되므로 부분 개체군 간에 의미 있는 개체 교환을 통해 효율성을 높이고, 더 좋은 성능을 기대할 수 있다. 분산 모델은 마이그레이션의 파라미터들을 어떻게 결정하는가에 따라 성능의 변화가 민감하게 나타날 수 있기 때문에 마이그레이션 간격, 이주자들의 개수, 이주자들의 선택 전략, 네트워크 토폴로지 등에 대한 결정이 이슈가 되고 있다. 셀 방식의 모델은 미립자 모델(Fine-Grained model)이라고도 한다. 이는 분산 모델과는 달리 개체군이 대량의 부분개체군으로 나뉘며, 각각의 프로세서에는 오직 하나의 개체가 할당된다. 선택 메카니즘과 교배 연산은 오직 이웃 개체에만 적용이 된다. 예를 들면 교배할 때, 개체의 이웃 개체들 가운데 가장 적합도가 높은 것과 교배를 하여 새로운 개체를 만들어낸다.

2.3 분산 유전 알고리즘

분산 유전 알고리즘의 기본적인 구조는 아래 [알고리즘 1] 과 같다[6]. 이전에 언급했듯이, DGA에서는 마이그레이션을 위한 파라미터들을 결정해야 하는데 핵심이 되는 파라미터들은 다음과 같다.

1) 마이그레이션 간격 : 마이그레이션 사이의 세대 간격을 의미한다.
2) 이주자들의 수 : 하나의 부분 개체군에서 다른 부분 개체군으로 몇 개의 개체들을 보내고 받을 것인지에 대한 사전 결정이 있어야한다.
3) 선택 전략 : 단일 유전 알고리즘에서는 일반적으로 선택이 적합도를 기준으로 정해진다. 분산 유전 알고리즘에서도 이주 시 선택 기준이 적합도가 사용될 수 있으나 다른 대안으로 다음과 같은 두 가지 방법이 많이 이용된다. 첫 번째 방법은 부분 개체군에서 개체를 무작위로 선택하여 마이그레이션 시키는 것이고, 두 번째 방법은 가장 좋은 개체를 이동시키는 방법이다. 전자는 성능이 좋지 않은 개체들에도 기회를 준다는 점에서 다양성을 높일 수 있다는 장점이 있는 반면, 다른 부분개체군으로 옮겨 갔을 때의 적응이 어려워 마이그레이션의 의미가 사라질 수 있고(Noneffect Problem), 가시적인 성능 향상이나 빠른 수렴보다는 장기적으로 최적해로의 수렴 가능성을 기대할 수 있다. 후자는 성능이 가장 좋은 개체를 마이그레이션함으로써 성능이 상대적으로 좋지 않은 부분개체군에서 가시적인 성능 향상과 빠른 수렴을 기대할 수 있지만 부분 개체군 간의 다양성을 잃을 수 있고, 모든 부분 개체군들이 전역적으로 우수한 개체의 영향을 강하게 받을 수 있다(Conquest Problem).
4) 마이그레이션 정책 : 마이그레이션 시 개체가 하나의 부분개체군에서 다른 부분개체군으로 이동할 것인지 아니면 복사될 것인지를 선택해야한다. 또한 이에 따라 부분개체군의 크기가 동적으로 변할 것인지 정적인

[Algorithm 1] Distributed Genetic Algorithm
1. Generate at random the population P of chromosomes.
2. Divide P into $SP_1, \ldots, SP_{N_s}$ subpopulations.
3. Define a neighborhood structure for SP_i, $i = 1, \ldots, N_s$
4. For SP_i, $i = 1, \ldots, N_s$
   Execute in parallel the next steps
      Apply, during f_m generations, the selection mechanism and the genetic operators.
      Calculate fitness.
      Send n_m chromosomes to neighboring subpopulations.
      Receive chromosomes from neighboring subpopulations.
   Until the stop criterion is satisfied.

크기를 유지할 것인지의 구체적인 구현 전략을 세워야 한다. 후자의 경우, 다른 부분개체군으로 보내진 개체들에 의해 원래 있던 개체들이 제거되어야 하므로 어떤 것을 제거할 것인지 결정해야한다. 임의로 제거하는 방법과 가장 적합도가 낮은 개체들을 제거하는 방법 등이 있다.

5) 네트워크 토폴로지 : 네트워크 토폴로지에서 섬(Island)의 이웃(Neighborhood)를 어떻게 정의하는지에 따라 두 가지 전략으로 나눌 수 있다. 징검돌(Stepping-stone) 방법은 모든 프로세서들 상호 간을 이웃으로 보기 때문에 자유롭게 마이그레이션 할 수 있다. 반면 섬 (Island) 방법은 지리적으로 가까운 프로세서 사이에만 마이그레이션 할 수 있다.

2.4 특징 선택

특징선택이란 어떤 학습(분류) 시스템 하에서 원본 데이터가 주어졌을 때, 가장 좋은 성능을 보여줄 수 있는 데이터의 부분집합(Subset)을 원본 데이터로부터 찾아내는 것을 말한다[8-9]. 특징부분집합을 평가하는 방법으로 독립적인 평가 함수를 사용하는 필터방식과 학습 알고리즘을 이용하는 래퍼방식이 있다.

3. 자동 마이그레이션 조절을 이용한 분산 유전 알고리즘

3.1 알고리즘 개요

단일 유전 알고리즘의 전체 개체군은 네트워크 토폴로지에 기반 한 여러 개의 부분 개체군들로 나뉜다. 각각의 부분개체군에서 유전 알고리즘은 독립적으로 수행된다. 매 세대에서 새롭게 탄생된 모든 개체들은 0살로 지정되며, 다음 세대에서 살아남을 때마다 1살씩 나이가 증가한다. 마이그레이션 간격으로 정의되는 고정된 수의 세대들이 지난 후, 임의로 선택된 부분개체군과 그것과 직접적으로 연결된 이웃들 사이에 개체들의 교환이 이루어진다. 마이그레이션이 될 개체들은 부분개체군들의 평균 진화 수준에 따라 자동적으로 선택된다. 부분개체군들의 진화 수준은 각각에 있는 개체들의 적합도의 평균에 의해 측정되며 부분개체군의 평균 적합도가 높을수록 상대적으로 진화 수준이 좋은 부분개체군으로 간주한다. 정복 문제를 피하고, 다양성을 유지하기 위해, 상대적으로 더 좋은 부분 개체군에 있는 월등히 좋은 적합도를 갖는 개체들은 더 낮은 부분 개체군으로 이주되지 않는다. 동시에, 타당하게 좋은 적합도를 갖는 개체들이 선택되어 진화에 기여하도록 하기 위해, 양방향으로 그들이 속하게 될 부분개체군의 평균 적합도보다 높으면서도 월등히 좋지는 않은 적합도를 갖는 개체들이 이주한다.

자세한 알고리즘은 아래의 알고리즘 2에 나와 있다.

자동 마이그레이션 조절 프로세스는 다음과 같다. 마이그레이션을 위해 선택된 두 개의 부분개체군을 각각 S1과 S2라 하자. 이들은 네트워크 토폴로지에서 상호 연결된 이웃

[Algorithm 2]
DGA using Automatic Control Migration

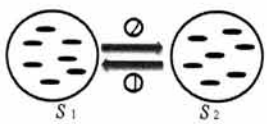
ivd_{ij} : j th individual in a subpopulation i
$A(ivd_{ij})$: age of ivd_{ij}
1. Generate at random the population P of individuals.
2. For all individuals ivd, $A(ivd)$ is initialized to 0.
3. Split P into $SP_1, \ldots, SP_{N_s}$ subpopulations
4. Define a neighborhood structure for
   SP_i, $i = 1, \ldots, N_s$
5. For SP_i, $i = 1, \ldots, N_s$
   Execute in parallel the next steps.
   5.1 During the fixed number of generations,
   Apply the selection mechanism and
   the genetic operators.
   Calculate fitness.
   If ivd_{ij} is generated by genetic operators,
   $A(ivd_{ij}) = 0$;
   Else
   $A(ivd_{ij})$++;

   5.2 Select one subpopulation SP_i randomly.
   For SP_i's neighbor sites,
   Perform migration process
   If ivd immigrates, $A(ivd_{ij}) = 0$;
   Until the stop criterion is satisfied.

관계에 있다. ivd_{ij}는 부분개체군 i에 있는 j번째 개체라 하자. $F(ivd_{ij})$가 ivd_{ij}의 적합도 라고 할 때,
부분개체군 i의 평균 적합도는 다음과 같다.

$$Avg\_F(S_i) = \frac{\sum_j F(ivd_{ij})}{|S_i|}$$

개체들은 양방향으로 이주한다. (그림 1)에서 각각 이주를 하는데 다른 기준이 적용됨을 보여준다. ①을 보면 개체들이 합류될 부분개체군의 평균보다 더 높은 적합도를 갖는다면 어떤 제약도 없이 이주된다. 반면 ②를 보면 이주하는 개체들은 합류할 부분개체군의 평균보다는 높지만 떠나는 부분개체군의 평균보다 더 낮은 적합도를 갖는다. 이로 인해 다양성을 유지하면서 동시에 정복문제의 가능성을 감소시킨다.



Assumption: $Avg\_F(S_1) > Avg\_F(S_2)$
① The directionof Migration: $S_2 -> S_1$
The Conditionof migrants: $\forall ivd_{2i}, F(ivd_{2i}) \geq Avg\_F(S_1)$
② The directionof Migration: $S_1 -> S_2$
The Conditionof migrants: $\forall ivd_{1i}, Avg\_F(S_2) \leq F(ivd_{1i}) \leq Avg\_F(S_1)$

(그림 1) 자동 마이그레이션 조절 프로세스

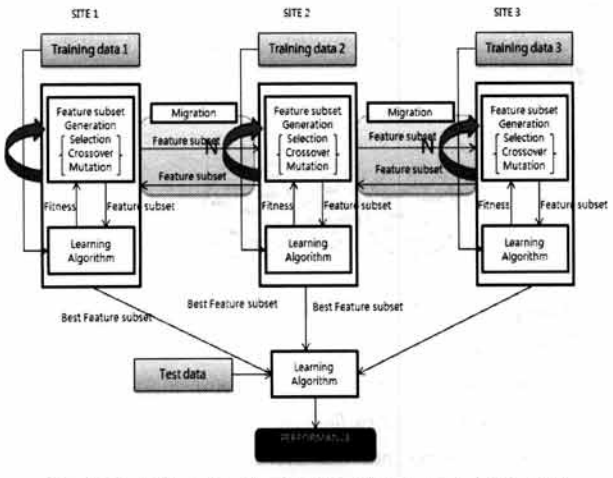
노화(Aging)에서 모든 개체들의 연대적 나이(Chronological Age)는 0으로 초기화된다. 만약 어떤 개체들이 다음 세대에도 선택되어 살아남는다면 그러한 개체들은 노화되어 나이는 1이 증가한다. 교배와 돌연변이에 의해 생겨난 개체의 나이는 0살이다. 마지막으로, 개체들이 다른 부분개체군들로 이동할 때, 그들의 나이는 새로운 환경에서 0부터 시작한다. 따라서 선택 메카니즘은 $A(ivd_{ij})$가 ivd_{ij}의 연대적 나이, I가 DGA의 반복횟수 그리고 α가 노화에 대한 상수 가중치라 할 때, ivd_{ij}의 선택 확률 $SP(ivd_{ij})$을 다음과 같이 정의한다.

$$SP(ivd_{ij}) = \alpha \times \frac{I - A(ivd_{ij})}{\sum_k (I - A(ivd_{ij}))} + (1-\alpha) \times \frac{F(ivd_{ij})}{\sum_k F(ivd_{ij})}$$

이러한 프로세스를 통해, 이주된 개체들은 다른 데이터 패턴으로 인해 새로운 부분개체군에서 낮은 적합도를 가질 수 있음에도 불구하고 어린 나이로 이러한 약점을 완화하고, 진화동안 그 개체들의 영향력 및 향후 생활을 성공적으로 보장해준다.

3.2 알고리즘을 이용한 특징선택

특징부분 선택은 전체 특징들 가운데 레이블 된 데이터에서의 분류 정확도(Classification Accuracy)와 특징 부분들과 결합되는 총비용과 같은 잘 알려진 기준에 대해 가장 좋은 성능을 보이는 최적의 특징들의 부분 집합을 찾는 것이다. 자동 마이그레이션을 적용한 DGA를 이용한 특징부분 선택은 학습 알고리즘을 이용한 래퍼방식을 따른다. 아래 (그림 2)는 분산 유전 알고리즘을 이용한 특징 부분 선택의 개요를 보여준다. 각각의 사이트에서 유전 알고리즘은 독립적으로 실행되며 산출된 특징부분들은 적합도 평가를 위해 학습 알고리즘으로 입력되고 훈련 데이터의 분류정확도를 계산한다. 제안된 자동 마이그레이션 조정에 의해 다음 세대의 개체군이 결정되고 이는 종료조건이 만족할 때 까지 반복된다. 최종 출력은 각각의 사이트에서 산출된 지역적으로 가장 좋은 부분집합들 사이에서 가장 좋은 특징 부분들이다.



(그림 2) 분산 유전 알고리즘을 이용한 특징 부분 선택

4. 실험 및 결과

4.1 실험방법

본 논문에서는 UCI machine learning archive로부터 6개의 데이터 셋들을 이용했다[11]. 각각의 데이터 셋에 있는 샘플들 중 30%는 테스트 데이터로서 이용되었으며 나머지는 학습 데이터로서 이용되었다. 분산된 환경을 위해 우리는 클래스들의 분포를 고려하여 각각의 트레이닝 데이터를 30개 또는 50개의 유사한 크기의 서브셋(Sub-set)으로 임의로 나눴다.

실험에서 특징 부분들의 적합도를 평가하고, 성능을 측정하기 위한 분류기로서 베이즈 이론(Bayes Theory)을바탕으로 한 나이브 베이즈분류기(Naïve Bayes Classifier, NB)[10]를 이용했다.

분류기는 공개 데이터 마이닝 실험 소프트웨어인 WEKA3.5 [12] 에 구현된 것을 사용하였다. 적합도 평가에서는, 트레이닝 데이터의 평균 분류 정확도를 계산하기 위해 10겹 교차검증 방식 (10-Fold Cross-Validation)을 사용하였다. 테스트 정확도(Test Accuracy)는 트레이닝 데이터에 속하지 않는 샘플들의 30%인 테스트 셋으로 계산된다.

4.2 데이터 셋과 파라미터 설명

실험에 사용된 6개의 데이터들에 대한 간략한 요약을 <표 1>에 정리했다.

〈표 1〉 데이터 셋

데이터셋	특징수	클래스수	샘플수	자식수	전체 인구수
Spambase	57	2	4601	50	500
Waveform	21	3	5000	50	500
Segment	19	7	2310	30	300
Splice	60	3	3190	50	500
Optdigits	64	10	3823	50	500
HDR Multifeature	649	10	2000	30	300

4.3 실험결과

실험에서 3.1 에 묘사된 것과 같은 특징 부분선택에 제안된 알고리즘을 적용하고 이를 단일 유전 알고리즘과 비교하였다. 실험에서 각각의 알고리즘은 30 세대동안 수행되었다. 분산 유전 알고리즘에서, 마이그레이션 간격은 5 세대로 했다.

<표 2>는 분류 정확도에 대한 CGA와 DGA의 성능을 보여준다. CGA에 대해서는, 훈련 정확도와 테스트 정확도를 볼 수 있고, DGA에 대해서는 다음 3가지 경우를 보여준다. 첫 번째 DGA는 우리의 자동조절 마이그레이션을 이용하였으나 선택 확률(Selection Probability)이 CGA처럼 적합도에 의해서만 결정된다. 나머지 두 개의 DGA는 본 논문에서 제안한 마이그레이션 과정과 선택 메카니즘이 모두 이용되었고 각각의 유존 알고리즘 선택 메카니즘에서 나이에 대한 가중치가 0.3과 0.7로 다른 조건으로 수행된다. 표에 있는 분

〈표 2〉 CGA와 DGA사이의 분류 정확도의 비교

데이터셋	CGA		DGA without aging	
	training	test	test	
			best	average
Spambase	89.88	89.67	89.29 (-)	82.68
Waveform	82.56	81.75	82.80 (+)	80.78
Segment	88.84	88.42	89.15 (+)	85.48
Splice	93.02	87.49	89.54 (+)	86.63
Optdigits	92.17	91.17	91.52 (+)	88.15
HDR Multi feature	96.46	92.67	92.51 (-)	91.59
Dataset	DGA using aging			
	age_weight=0.3		age_weight=0.7	
	test		test	
	best	average	best	average
Spambase	88.85 (-)	82.19	89.07 (-)	83.13
Waveform	82.34 (+)	80.86	82.67 (+)	81.17
Segment	89.59 (+)	85.65	89.15 (+)	86.60
Splice	90.56 (+)	87.11	90.15 (+)	86.57
Optdigits	90.92 (-)	87.87	90.83 (-)	88.31
HDR Multi feature	92.67 (~)	91.59	92.67 (~)	91.66

산 유전 알고리즘에서 best는 모든 부분 개체군들의 테스트 정확도들 중 가장 좋은 것을 나타내며, average는 그것들의 평균 테스트 정확도를 의미한다. <표 2>에서 CGA와 비교했을 때, DGA가 더 좋은 성능을 보이는 경우, 즉 테스트 정확도가 더 좋은 경우에 (+) 기호를 사용하였으며, 반대의 경우 (-) 기호를, 같거나 비슷한 경우 (~) 기호를 사용하였다. 또한 가장 좋은 성능을 보이는 경우 정확도를 굵은 글씨로 나타내었다.

실험 결과 DGA의 대부분이 주어진 데이터셋에 대해 CGA보다 좋은 성능을 낸다. 그러나 DGA와 CGA의 성능 차이는 2% 이내이므로 근소하다. 그렇지만 CGA와 비교하였을 때, 병렬 계산으로 인한 연산시간 등 의 이점을 고려하면 DGA를 이용하는 것이 유리함을 알 수 있다. 게다가 <표 3>을 보면 모든 지역적으로 가장 좋은 개체들이 작은 표준 편차로 일정한 성능을 산출한다. 특히 age_weight를

〈표 3〉 각 테스트 정확도 표준편차

데이터셋	표준편차		
	DGA without aging	DGA using aging	
		age_weight=0.3	age_weight=0.7
Spambase	5.10	6.38	4.83
Waveform	1.26	1.29	0.95
segment	3.49	3.44	1.65
Splice	2.15	1.66	2.21
Optdigits	1.79	2.02	1.75
HDR Multi feature	0.56	0.58	0.48

0.7로 준 결과를 보면 Spambase를 제외하고 나머지는 작은 표준편차를 보였다. 이것은 연대적 나이의 가중치를 크게 줌으로써, 부분 개체군들 간에 마이그레이션 된 개체가 도태되지 않도록 하고, 각각의 부분개체군에서 좋은 개체라고 하더라도 오랜 세대동안 진화에 상당한 영향을 준 이후에는 연대적 나이로 인해 선택확률이 낮아지므로 좋은 개체와 나쁜 개체들이 적절히 섞여 다양성을 유지할 수 있는 결과라고 보여 진다. 병렬화로 인한 네트워크 오버헤드는 각 사이트에서 각각 유전 알고리즘이 병렬적으로 분산되어 실행되기 때문에 병렬화로 인한 오버헤드는 크게 발생하지 않는다. 앞의 2장에서 소개 한 바와 같이 기존 마이그레이션 선택전략과 달리 본 논문에서 제시하는 방법은 이동한 개체가 그곳에서 의미가 없어지는 문제와 우수한 개체가 나머지를 지배하는 정복문제를 모두 고려한 방법이다. 각 사이트의 평균값을 기준으로 이주를 결정하기 때문에 이주하는 개체의 수도 자동으로 결정이 된다. 우수한 개체가 이동하는 방법을 사용할 때 이주자의 수를 결정하는 문제 또한 성능에 큰 영향을 미치는데 이를 지능적으로 해결할 수 있다. 이렇게 평균값을 기준으로 양방향으로 이동하기 때문에 네트워크의 설계는 섬 토폴로지(Topology)를 사용했다. 일반적으로 분산된 모든 사이트를 연결(Fully Connected)하는 방법은 중앙에서 모든 사이트를 관장하는 사이트를 두어야 하며 설계의 비용이 비싸나 성능은 우수하다. 제안된 섬 토폴로지를 사용한 이유는 주위 이웃만을 고려하기 때문에 시간적인 이점이 있고 중앙에 관리 사이트를 설계하지 않아도 되어 비용이 절감된다. 그러나 이동이 제한되어 성능이 떨어질 수 있으나 제안된 자동 마이그레이션 알고리즘이 양방향의 이동을 고려하기 때문에 단순한 섬 토폴로지를 사용했다. 이는 제안된 알고리즘의 성능이 각 사이트에서 고르게 수행됨을 표준편차를 이용해 보였다. 이는 도태되는 사이트 없이 각 사이트에서 학습된 결과의 질(Quality)이 우수함을 보여준다.

<표 4>는 기존 DGA와 제안된 DGA의 정확도를 비교한다. 자동화된 마이그레이션 방법을 사용했을 때 성능의 개선이 있음을 볼 수 있다.

〈표 4〉 기존 DGA와 제안된 DGA 정확도 비교

데이터셋	기존 DGA		제안된 DGA	
	best	average	best	average
Spambase	89.07	82.34	89.29 (+)	82.68
Waveform	82.74	80.85	82.80 (+)	80.78
Segment	89.88	85.81	89.15 (-)	85.48
Splice	89.23	86.98	89.54 (+)	86.63
Optdigits	91.35	87.63	91.52 (+)	88.15
HDR Multifeature	92.51	91.60	92.51 (~)	91.59

5. 결론 및 향후과제

본 논문은 분산된 데이터를 병렬로 수행하는 유전 알고리

즘을 제안하였다. 제안된 알고리즘은 각각의 부분개체군에 있는 개체들이 사용자 지정 파라미터에 의존하지 않고 다른 사이트로 이동하고 개체군과 결합한다. 결과적으로 개체군들 사이의 다양성 뿐 아니라 모든 개체군들에서 적합도의 연속적인 개선을 돕는다. 본 논문에서 제안한 알고리즘이 기존의 분산 유전 알고리즘과 비교하여 개선된 점은 다음과 같다.

먼저 마이그레이션에 따른 파라미터의 결정 문제를 자동적으로 해결할 수 있다. 제안된 알고리즘은 부분개체군들의 질에 따라 적절하게 마이그레이션의 주요 파라미터인 이주자의 수와 이주자들의 선택 문제를 결정해준다. 그리고 양방향 마이그레이션 전략을 수행하며, 각 방향에 대해 이주자의 선택에 서로 다른 기준을 적용함으로써 부분개체군들의 균형 있는 진화와 성능 향상을 돕고, 기존 분산 유전 알고리즘들이 빠지기 쉬운 정복 문제의 위험을 감소시키는 기준을 적용함에 따라 각각의 부분개체군들의 다양성을 유지하도록 시도한다. 마지막으로 노화의 개념을 이용한 선택 전략은 마이그레이션이 된 개체들의 향후 생활을 도움으로써 이주자들이 새로운 데이터 패턴에도 도태되지 않고 일정 기간 진화에 영향을 줄 수 있도록 보장해준다.

본 논문에서 앞으로 연구해야할 향후 연구과제로는 현재 각 부분개체군들이 같은 연산들을 사용하는 동질(Homogeneous) DGA를 이용하였지만, 서로 다른 연산들을 사용하는 이질(Heterogeneous) DGA를 이용하는 것과 실험에서 마이그레이션 구간과 노화 가중치와 같은 파라미터 셋팅 시 한정된 파라미터 셋팅을 하나 추가적인 데이터 셋에서 다른 파라미터 셋팅을 이용한 확장된 실험을 할 수 있을 것이다. 그리고 데이터의 양에 따른 적절한 네트워크 토폴로지에 대해 고려해 볼 수 있다.

참 고 문 헌

[1] H. Kargupta, and P. Chan, "Advancedin Distributed and Parallel Knowledge Discovery," AAAIPress/MITPress, California, 2000.

[2] J. H. Holland, "Adaptation in Natural and Artificial systems," AnnArbor : The University of Michigan Press, Michigan, 1975.

[3] D. E. Goldberg, "Genetic algorithms in Search, Optimization and Machine Learning," Addison-Wesley Longman, Inc, Boston, 1989.

[4] D. V. Michael, "The Simple Genetic Algorithm : Foundations and Theory," MITPress, California, 1999.

[5] E. Alba, "Parallel Metaheuristics : A New Class of Algorithms," JohnWiley & Sons, Inc., New Jersey, 2005.

[6] F. Herrera and M. Lozano, "Gradual distributed real-coded genetic algorithms," IEEE Transactions on Evolutionary Computation, 2000, Vol.4, No.1, pp.43-63.

[7] E. Cant-paz, "A survey of parallel genetic algorithm," Calculateurs Paralleles, 1998, Vol.10, pp.141-171.

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

[9] J. Yang and V. Honavar, "Feature subset selection using a Genetic Algorithm," IEEE Intelligent Systems, 1998, Vol.13, No.2, pp.44-49.

[10] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss", Machine Learning, 1997, Vol. 29, pp. 103-137.

[11] A. Asuncion and D. Newman, UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, School of Information and Computer Science, 2007.

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

이 현 정

e-mail : luckyhj777@naver.com
2007년 서강대학교 컴퓨터공학과(학사)
2009년 서강대학교 컴퓨터공학과(석사)
2009년~현 재 삼성전자 DMC 연구소 재직
관심분야 : 유전 알고리즘, 분산 알고리즘

나 용 찬

e-mail : ycna@sogang.ac.kr
1999년 단국대학교 컴퓨터공학과(학사)
2002년 단국대학교 컴퓨터공학과(석사)
2005년 서강대학교 컴퓨터공학과 박사과정
관심분야 : 베이지안 네트워크 구조학습, 분산 데이터 학습 등

양 지 훈

e-mail : yangjh@sogang.ac.kr
1987년 서강대학교 전자계산학과(학사)
1989년 ISU Department of Computer Science (석사)
1999년 ISU Department of Computer Science (박사)
1999년~2000년 HRL Laboratories, LLC, Malibu, CA 연구원
2000년~2002년 SRA International, Inco, Fairfax, VA 연구원
2002년~현 재 서강대학교 컴퓨터공학과 부교수
관심분야 : 기계학습, 바이오인포메틱스 등