유전학 기반 학습 환경하에서 분류 시스템의 성능 향상을 위한 엔-버전 학습법

김 영 준† · 홍 철 의†

요 약

델보는 주어진 사례의 집합으로부터 이들 사례들을 분류할 수 있는 베이지안 분류 규칙들로 이루어진 규칙 집합을 습득하는 유전학 기반 귀납적 학습 시스템이다. 규칙 집합의 습득과정에서 델보가 당면하게 되는 한 가지 문제점은 학습 과정이 최적의 규칙 집합이 아닌 지역 최적치를 습득하고 종료하는 경우가 가끔 발생한다는 것이다. 다른 하나의 문제점은 훈련 사례에 대한 경우와는 달리 새로운 평가 사례에 대해 분류 성능이 현저히 저하되는 규칙 집합을 습득하는 경우가 가끔 발생한다는 것이다. 본 논문에서는 이러한 문제점을 해결하여 보다 성능이 향상된 분류 시스템을 구축하기 위한 기법으로 엔-버전 학습법에 관해 연구하였다. 엔-버전 학습법은 주어진 사례 집합에 대해 다수의 규칙 집합을 습득한 후 이를 이용하여 분류 시스템을 구축함으로써 분류 시스템의 전체적인 성능을 향상시키는 기법이다. 엔-버전 학습법의 구현을 위해 다수의 규칙 집합을 이용하여 최종 분류 결과를 도출해 내기 위한 기법과 습득된 규칙 집합들로부터 분류 시스템을 구축하기 위한 최적의 규칙 집합의 조합을 찾기 위한 기법을 제시하고 다수의 사례 집합을 이용하여 엔-버전 학습법이 델보의 학습 환경에 미치는 영향을 평가하였다.

An N-version Learning Approach to Enhance the Prediction Accuracy of Classification Systems in Genetics-based Learning Environments

Yeong-Joon Kim† · Chul-Eui Hong†

ABSTRACT

DELVAUX is a genetics-based inductive learning system that learns a rule-set, which consists of Bayesian classification rules, from sets of examples for classification tasks. One problem that DELVAUX faces in the rule-set learning process is that, occasionally, the learning process ends with a local optimum without finding the best rule-set. Another problem is that, occasionally, the learning process ends with a rule-set that performs well for the training examples but not for the unknown testing examples. This paper describes efforts to alleviate these two problems centering on the *N-version learning approach*, in which multiple rule-sets are learned and a classification system is constructed with those learned rule-sets to improve the overall performance of a classification system. For the implementation of the N-version learning approach, we propose a decision-making scheme that can draw a decision using multiple rule-sets and a genetic algorithm approach to find a good combination of rule-sets from a set of learned rule-sets. We also present empirical results that evaluate the effect of the N-version learning approach in the DELVAUX learning environment.

† 정 회 원 : 상명대학교 정보통신학부 교수
논문접수 : 1999년 1월 13일, 심사완료 : 1999년 6월 7일

1. 서 론

유전자 알고리즘은 주어진 문제에 대하여 이진 문자를 이용하여 코딩 된 가능한 해들로 개체 집단을 생성한 후 개체 집단내의 구성원에 생물학적 진화 과정에서 볼 수 있는 유전 연산자들을 적용하여 새로운 개체 집단을 생성하는 과정을 반복하면서 주어진 문제의 최적 해를 찾는 탐색 알고리즘이다. 유전자 알고리즘은 일반적인 탐색 문제 및 여러 최적화 문제 등의 해결에 널리 이용되어 왔다[1,2].

델보(DELVAUX)는 유전자 알고리즘을 이용하여 주어진 사례의 집합으로부터 이들 사례들을 분류할 수 있는 베이지안 분류 규칙들로 이루어진 규칙 집합을 습득하는 귀납적 학습 시스템이다[3,4]. 유전자 알고리즘을 이용한 학습 시스템의 구축에 관한 연구는 이미 오래 전부터 진행되어 왔으나 대부분의 연구는 참과 거짓의 두 진리 값을 갖는 생성 규칙들을 습득하는 학습 시스템의 구축에 관한 것이었으며[5,6,7,8], 최근에는 퍼지 컨트롤러의 구현에 필요한 퍼지 규칙들을 습득하기 위한 학습 시스템의 구현에 유전자 알고리즘을 이용하기 위한 연구[9,10,11]와 퍼지 규칙에 기반을 둔 분류 시스템의 구축에 유전자 알고리즘을 이용하기 위한 연구[12,13,14]가 활발히 진행되고 있다. 그러나 델보는 불확실성 하에서 통계적, 확률적 정보에 준하여 추론이 가능한 다치 값에 기반을 둔 베이지안 분류 규칙들을 습득하기 위한 학습 시스템으로, 기존의 참과 거짓의 두 진리 값에 기반을 둔 생성 규칙을 습득하기 위한 학습 시스템이나 동적인 환경 하에서 시스템을 통제하기에 적합한 퍼지 규칙을 습득하기 위한 학습 시스템과는 다소 차이가 있다고 하겠다.

델보를 이용한 분류 규칙 집합의 습득 시 당면하게 되는 한 가지 문제점은 학습 시스템이 주어진 사례 집합에 대해 최적의 규칙 집합을 습득하지 못하여 학습 결과로 구축된 분류 시스템의 성능이 기대에 못 미치는 결과를 초래하는 경우가 가끔 발생한다는 것이다. 다른 하나의 문제점은 습득된 분류 규칙 집합이 새로운 사례 집합에 대해 학습 과정에서 보여준 분류 성능을 제공하지 못하는 경우가 가끔 발생하는 것이다. 본 논문에서는 델보를 이용한 분류 시스템의 구축 시 당면하게 되는 이러한 문제점을 해결하여 보다 성능이 향상된 분류 시스템을 구축하기 위한 기법으로 엔-버전 학습법에 관해 연구하였다. 엔-버전 학습법은 주어진 사례 집합에 대해 다수의 규칙 집합을 습득한 후 이를 이용하여 분류 시스템을 구축함으로써 분류 시스템의 전체적인 성능을 향상시키는 기법이다.

본 논문의 구성은 다음과 같다. 2장에서는 델보 시스템을 베이지안 분류 규칙의 구문 형태 및 의미, 유전자 알고리즘을 이용한 학습 시스템의 구현을 중심으로 간략히 소개한다. 3장에서는 분류 시스템의 성능 향상을 위한 엔-버전 학습법에 대해 자세히 설명하고, 4장에서는 엔-버전 학습법이 유전자 학습 환경의 학습 능력 향상에 미치는 영향을 다양한 사례 집합을 이용하여 평가하였다. 5장은 결론 및 향후 과제에 대해 설명한다.

2. 델보의 학습 환경

델보는 유전자 알고리즘을 이용하여 주어진 사례의 집합으로부터 이들 사례들을 분류할 수 있는 분류 규칙 집합을 습득하는 귀납적 학습 시스템이다. 본 장에서는 델보가 습득하게 되는 분류 규칙의 구문 형태 및 의미, 유전자 알고리즘을 이용한 학습 시스템의 구현을 중심으로 델보 시스템을 간략히 소개 한다.

2.1 베이지안 분류 규칙

델보의 학습 환경에서 훈련 사례 집합내의 각각의 사례는 속성 A_1, A_2, ..., A_n에 대한 값 a_1, a_2, ..., a_n과 그 사례가 속한 클래스 c로 구성된 리스트, (a_1, a_2, ..., a_n, c)의 형태로 표현된다. 각 사례 e의 속성 A_i에 대한 값 a_i는 사례 집합 내에서 사례 e의 속성 A_i에 대한 실제 속성 값보다 적은 값을 갖는 사례의 수를 사례 e의 실제 속성 값과 다른 값을 가지는 사례의 수로 나누어 0과 1사이의 값을 갖도록 정규화한 값이다.

학습 시스템은 주어진 사례의 집합으로부터 "If E then C with S = s, N =n" 형태의 분류 규칙들을 습득한다. 이들 습득된 분류 규칙으로 구축된 분류 시스템에서는 주어진 사례가 어떤 특정 클래스에 C에 속할 사전 가능성 O(C)를 그 사례가 클래스 C에 속할 확률 P(C)로부터 $O(C) = P(C) / (1 - P(C))$의 식을 이용하여 구한다. 클래스 C에 속할 확률 P(C)는 주어진 사례의 집합 내에서 클래스 C에 속하는 사례가 차지하는 비율로부터 구한다. 주어진 사례에 대해 각각의 규칙들은 분류 규칙의 조건(즉, E)에서 고려하는 속성에 대해 그 사례가 갖고있는 속성 값에 따라 주어진 사례가 분류

규칙의 결론인 클래스 C에 속할 가능성에 대한 승수를 N과 S사이의 값으로 제공한다. 각각의 분류 규칙은 주어진 사례가 분류 규칙의 조건을 완전하게 만족하면(즉, P(E′) = 1) S의 값을, 불만족 시에는(즉, P(E′) = 0) N의 값을, 0 < P(E′) < 1인 경우에는 P(E′)의 값에 비례하여 N과 S사이의 값을 제공한다. 분류 시스템은 주어진 사례에 대해 그 사례가 특정 클래스 C에 속할 사전 가능성 O(C)에 클래스 C를 결론으로 갖는 분류 규칙들이 제공하는 가능성에 대한 승수를 베이지안 추론 법에 준하여 취합하여 클래스 C에 속할 사후 가능성 O(C′)을 구한다. 분류 시스템은 각각의 클래스에 대해 주어진 사례가 그 클래스에 속할 사후 가능성을 구한 후 사후 가능성이 가장 큰 클래스를 주어진 사례가 속한 클래스로 선택한다.

학습 시스템이 주어진 사례의 집합으로부터 습득하는 분류 규칙의 형태 중 하나는 "If is-high(A) then C with S = s, N = n"의 형태로 이 타입의 규칙은 고려대상이 되는 속성 A의 값의 상대적인 높고 낮음에 따라 주어진 사례가 클래스 C에 속할 가능성에 대한 승수를 N과 S사이의 값으로 제공한다. 다른 하나는 "If is-close(A, a) then C with S = s, N= n"의 형태로 고려하는 속성 A의 값이 어떤 특정 값 a에 근사한 정도에 따라 클래스 C에 대해 N과 S사이의 값을 제공한다. 학습 시스템은 주어진 사례의 집합에 대해 그 사례들을 분류하기 위한 규칙 집합(즉, 주어진 사례들을 분류하기 위해 필요한 속성들을 적절히 고려한 is-high 규칙과 is-close 규칙들)을 각각의 규칙에 필요한 s, n, 상수 a의 값(is-close 규칙에 한 함)과 함께 유전자 알고리즘을 이용하여 습득하는 것이다. (그림 1)은 4개의 속성 A_1, A_2, A_3, A_4를 갖고 클래스 C_1, C_2, C_3 중 하나의 클래스에 속하는 사례들로 구성된 훈련 사례 집합을 이용하여 습득된 분류 규칙 집합의 예를 보인 것이다.

If is-high(A_1)
  then C_1 with S=1944, N=0.016
If is-high(A_3)
  then C_2 with S=0.002, N=3139980
If is-close(A_4,0.0198)
  then C_2 with S=0.8, N=534
If is-high(A_3)
  then C_1 with S=0.42, N=4.9e+12
If is-close(A_4,0.774)
  then C_2 with S=3.2e+6, N= 0.003
If is-high(A_4)
  then C_3 with S=9.9e+18, N=0.21

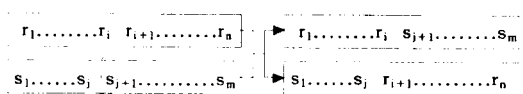
(그림 1) 델보에 의해 습득된 분류 규칙 집합의 예

2.2 유전자 알고리즘을 이용한 학습 시스템의 구현

유전자 알고리즘을 이용한 학습 시스템의 구현에서 개체 집단은 일정 수의 분류 규칙 집합으로 구성되며, 각각의 분류 규칙 집합은 임의의 수의 베이지안 분류 규칙으로 구성된다. 즉, 개체 집단의 구성원은 임의의 수의 베이지안 분류 규칙들로 이루어진 분류 규칙의 집합이다.

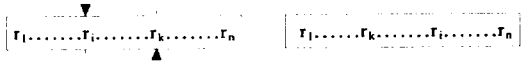
초기의 개체 집단은 난수 발생기를 이용하여 임의의 수의 분류 규칙들로 이루어진 일정 수의 분류 규칙 집합을 생성함으로써 얻어진다. 계속되는 진화 과정에서는 적합도에 비례하여 선택된 분류 규칙 집합에 대해 유전 연산자를 적용하여 새로운 개체 집단을 생성하는 과정을 반복한다. 각각의 분류 규칙 집합의 적합도는 분류 규칙 집합이 주어진 사례 집합내의 사례들을 어느 정도 정확하게 분류 할 수 있는가 하는 분류의 정확도를 이용하여 평가한다. 주어진 개체 집단에서 새로운 개체 집단을 생성하는 과정을 원하는 해가 얻어질 때까지 반복한다.

주어진 개체 집단에서 새로운 개체 집단을 생성하는 과정에서 교배 연산자, 돌연변이 연산자, 역치 연산자가 이용되었다. 교배 연산자는 선택된 두 개의 규칙 집합내의 규칙들의 일부를 교환하여 새로운 규칙 집합을 생성한다((그림 2) (a)). 돌연 변이 연산자는 선택된 규칙 집합내의 하나의 규칙을 선택하여 새로운 규칙으로 대체함으로써 새로운 규칙 집합을 생성한다((그림 2) (b)). 역치 연산자는 규칙 집합내의 두 개의 규칙을 선택하여 이들의 위치를 교환해 줌으로써 새로운 규칙 집합을 생성한다((그림 2) (c)). 역치 연산자의 역할은 서로 연관성이 있는 분류 규칙들을 인접한 위치에 놓이게 함으로써 교배 연산자를 이용하여 새로운 규칙 집합을 생성할 때 이들 연관된 규칙들이 분리 될 가능성을 감소시켜 가능한 한 같은 분류 규칙 집합 내에 있도록 하기 위해 사용되었다.



(a) 교배 연산자

(b) 돌연변이 연산자



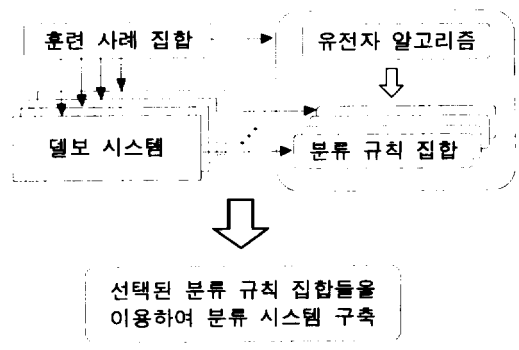
(c) 역치 연산자

(그림 2) 델보의 유전 연산자

3. 엔-버전 학습법

주어진 사례 집합에 대한 최적의 분류 규칙 집합을 습득하기 위한 학습 과정의 초기 단계에서 델보는 난수 발생기를 이용하여 일정 수의 규칙 집합으로 구성된 개체 집단을 생성한다. 계속되는 학습 과정에서는 개체 집단내의 구성원에 유전 연산자들을 적용하여 새로운 개체 집단을 생성하는 과정을 만족할 만한 규칙 집합을 습득할 때까지 반복하는데 이 과정에서 유전 연산자를 적용하기 위한 규칙 집합의 선택, 임의의 규칙을 다른 새로운 규칙으로 교체하기 위한 규칙의 선택 및 규칙의 생성 등에 난수 발생기가 이용된다. 델보의 난수 발생기에 의존한 학습 과정은 난수 발생기에 사용되는 초기값이 변함에 따라 다른 탐색 공간을 탐색하여 결과적으로 다른 분류 규칙 집합을 습득하게 되는데 델보의 이러한 특성을 이용하여 엔-버전 학습법을 델보의 학습 환경에서 구현하였다. 엔-버전 학습법은 주어진 사례 집합에 대해 다수의 분류 규칙 집합을 습득한 후 이를 이용하여 분류 시스템을 구축함으로써 분류 시스템의 전체적인 성능을 향상시키는 기법이다.

델보의 학습 환경하에서 엔-버전 학습법의 구현은

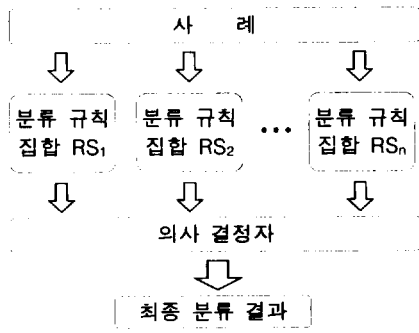


(그림 3) 유전학 기반 학습 환경에서의 엔-버전 학습법

두 단계로 이루어진다. 첫 단계는 다수의 분류 규칙 집합을 습득하기 위한 단계로 이 단계에서는 주어진 사례 집합에 대해 델보를 반복 실행시켜 다수의 규칙 집합을 습득한다. 두 번째 단계에서는 이들 습득된 규칙 집합들에 대해 유전자 알고리즘을 적용하여 최적의 규칙 집합의 조합을 얻어낸 후 이를 이용하여 분류 시스템을 구축한다. (그림 3)은 엔-버전 학습법을 나타낸 것이다.

3.1 엔-버전 학습법을 위한 의사 결정 기법

(그림 4)는 엔-버전 학습법 하에서 구축한 분류 시스템의 구조를 보인 것이다. 분류 시스템은 다수의 분류 규칙 집합과 의사 결정자로 구성된다. 사례에 대한 분류 과정에서 분류 규칙 집합은 최종 분류 결과를 도출해 내기 위해 필요한 자료를 의사 결정자에게 제공하고 의사 결정자는 이들 자료를 적절한 방법에 따라 취합하여 주어진 사례가 속할 클래스를 결정한다. 따라서 엔-버전 학습법의 구현을 위해서는 분류 규칙 집합이 제공하는 자료를 취합하여 최종 결론을 도출해 내기 위한 의사 결정 기법이 필요하다.



(그림 4) 엔-버전 학습법 하에서 구축된 분류 시스템의 구조

엔-버전 학습법의 구현을 위해 이용 가능한 의사 결정 기법 중 하나는 단순 보우팅(voting)을 이용하는 것이다. 이 기법하에서 각각의 분류 규칙 집합은 2.1절에서 논한 방법에 따라 사후 가능성이 가장 큰 클래스를 정하여 그 결과를 의사 결정자에게 제공하고, 의사 결정자는 다수결의 원칙에 따라 주어진 사례가 속할 클래스를 최종적으로 선택한다. 그러나 이 단순 보우팅 기법은 구현이 간단한 반면에 사례가 다른 클래스에 속할 사후 가능성을 의사 결정과정에 반영하지 못하는 단점이 있다.

본 논문에서는 단순 보우팅 기법보다는 좀 더 정확한 분류 결과를 얻을 수 있으리라는 기대하에 분류 규칙 집합이 제공하는 정규화된 사후 가능성을 취합한 후 그 결과에 준하여 사례가 속할 클래스를 선택하는 기법을 개발하였다. 이 기법하에서 각각의 분류 규칙 집합은 2.1절에서 논한 방법으로 구한 사후 가능성들을 식 (1)에 따라 정규화하여 그 결과를 의사 결정자에게 제공한다.

$$NO(C_j') = \frac{O(C_j')}{\max_k\{O(C_k')\}} \quad (1)$$

식 (1)에서 $NO(C_j')$, $O(C_j')$는 각각 클래스 C_j에 대한 정규화된 사후 가능성 및 사후 가능성을 나타내고, $max_k\{O(C_k')\}$는 클래스들에 대한 사후 가능성 중 가장 큰 값을 반환하는 함수이다. 의사 결정자는 정규화된 사후 가능성들을 각각의 클래스 C_j에 대해 식 (2)에 따라 취합한 후 CE 값이 가장 큰 클래스를 주어진 사례가 속한 클래스로 선택한다.

$$CE(C_j') = \left(\prod_{i=1}^{n}(1 + NO_{RS_i}(C_j'))\right)^{\frac{1}{n}} \quad (2)$$

식 (2)에서 $NO_{RSi}(C_j')$는 분류 규칙 집합 RS_i가 클래스 C_j에 대해 제공하는 정규화된 사후 가능성을, n은 분류 시스템내의 분류 규칙 집합의 수를 나타낸다.

식 (1)을 이용한 정규화 과정은 한 분류 규칙 집합의 사후 가능성이 다른 분류 규칙 집합의 사후 가능성에 비해 지나치게 큰 경우로 인해 의사 결정 과정에 절대적인 영향을 미치지 않도록 분류 규칙 집합들 사이에 존재하는 사후 가능성에 대한 편차를 줄이기 위한 조치이다. 식 (2)에서는 0에 근접한 정규화된 사후 가능성이 전체 의사 결정에 영향을 미치지 않도록 하기 위해 정규화된 사후 가능성에 1의 값을 더하였다.

3.2 최적 분류 규칙 집합의 조합 습득을 위한 유전자 알고리즘

엔-버전 학습법의 두 번째 단계는 전 단계에서 습득된 다수의 분류 규칙 집합에 탐색 알고리즘을 적용하여 최적의 분류 규칙 집합의 조합을 찾아내는 것이다. 본 연구에서는 엔-버전 학습법의 두 번째 단계를 유전자 알고리즘을 이용하여 구현하였다.

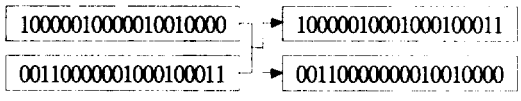
유전자 알고리즘을 이용한 구현에서 개체 집단의 구성원은 엔-버전 학습법의 첫 단계에서 습득한 규칙 집합의 수 만큼의 이진 문자로 표현되며, 각각의 이진 문자는 상응하는 규칙 집합이 구성원에 포함되는지의 여부에 따라 0과 1로 표현된다. 즉, 개체 집단의 각각의 구성원은 엔-버전 학습법의 첫 단계에서 습득된 규칙 집합들 중의 일부로 구성된다. (그림 5)는 엔-버전 학습법의 첫 단계에서 20개의 분류 규칙 집합 RS_1, RS_2, ..., RS_{20}을 습득하였다는 가정 하에 개체 집단 구성원의 예를 보인 것이다.

{RS_1,RS_7,RS_{13},RS_{16}} : 1000001000001010000

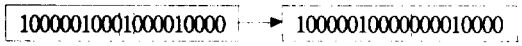
{RS_3,RS_4,RS_{11},RS_{15},RS_{20}} : 00110000001000100001

(그림 5) 엔-버전 학습법의 두 번째 단계를 위한 구성원의 염색체적 표현

규칙 집합의 조합 탐색을 위한 초기 단계에서는 일정 수의 구성원을 갖는 개체 집단을 난수 발생기를 이용하여 생성한 후 계속되는 진화 과정에서는 개체 집단내의 구성원에 교배 연산자, 돌연변이 연산자를 적용하여 새로운 개체 집단을 생성하는 과정을 만족할 만한 분류 규칙 집합의 조합을 습득할 때까지 반복한다. 교배 연산자는 현재의 개체 집단에서 2개의 구성원을 선택하여 구성원내의 규칙 집합의 일부를 교환하여 새로운 개체를 생성한다((그림 6) (a)). 돌연 변이 연산자는 선택된 구성원 내의 이진 문자 하나를 다른 문자로 교체하여 규칙 집합을 제거하거나 추가하는 과정을 통해 새로운 개체를 생성한다((그림 6) (b)). 각각의 구성원의 적합도는 구성원 내의 분류 규칙 집합들로 3.1절에서 언급한 방법에 따라 분류 시스템을 구축한 후 주어진 훈련 사례 집합을 어느 정도 정확하게 분류해 내느냐 하는 분류의 정확도로 평가한다.



(a) 교배 연산자



(b) 돌연변이 연산자

(그림 6) 엔-버전 학습법의 두 번째 단계를 위한 유전 연산자

(그림 7)은 유전자 알고리즘을 이용한 구현에서 주어진 개체 집단 G(t)로부터 새로운 개체 집단 G(t+1)을 생성하는 과정을 보인 것이다. (그림 7)에서 함수 *best...*는 현재의 개체 집단에서 최고의 적합도를 갖는 구성원을 다음 개체 집단에 포함시키기 위한 함수이다. *should-...* 함수는 돌연 변이와 교배의 확률에 따라 참과 거짓을 반환하는 함수로 선택된 구성원에 돌연 변이나 교배 연산자를 주어진 확률에 따라 적용하기 위해 사용한다. *crossover* 함수는 선택된 두 개의 구성원에 교배 연산자를 적용하여 새로운 구성원을 생성한 후 이를 반환하는 함수이다. *mutate*는 선택된 규칙 집합에 돌연변이 연산자를 적용하여 그 결과를 반환한다.

```
next-generation(t):=
Insert best-member-of(G(t)) into G(t+1);
DO { Select two members M1 and M2( M1 ≠ M2)
     using roulette-wheel;
     If should-make-crossover()
       then crossover(M1, M2, NM1, NM2)
       else NM1 := M1 ; NM2 := M2;
     If should-have-mutation()
       then NM1 := mutate(NM1);
     If should-have-mutation()
       then NM2 := mutate(NM2);
     Insert NM1, NM2 into G(t+1); }
UNTIL population of G(t+1) is complete;
```

(그림 7) 새로운 개체 집단의 생성

4. 엔-버전 학습법이 유전자 학습 환경에 미치는 영향의 평가

엔-버전 학습법이 유전자 학습 환경의 학습 능력 향상에 미치는 영향을 다양한 사례 집합을 이용하여 평가하였다. 엔-버전 학습법의 평가를 위해 사용한 사례 집합은 다음과 같다[1].

● 붓꽃 사례 집합 : 3가지 종류의 붓꽃으로부터 얻어진 150개(각 종류별 50개 씩)의 사례 집합으로 각각의 사례는 꽃잎의 길이, 꽃잎의 넓이, 꽃받침의 길이, 꽃받침의 넓이의 4가지 속성 값과 그 사례가 속한 붓꽃의 종류를 나타내는 값으로 표현된다.

● 유리 사례 집합 : 6 종류의 유리 파편들로부터 얻어진 214개의 사례로 구성되어 있으며, 각각의 사례는 유리를 구성하는 9가지 물질의 성분비 및 그 사례가 속한 유리의 종류를 나타내는 값으로 표현된 사례 집합이다.

● 당뇨 환자 사례 집합 : 당뇨 환자인 경우와 정상인인 경우로 분류되는 768개의 사례로 구성되었으며, 각각의 사례는 8가지 속성 값과 그 사례가 속한 클래스로 표현된다.

● 레이다 시그널 사례 집합 : 올바른 경우와 잘못된 경우의 351개 레이다 시그널 사례로 구성되었으며, 각각의 사례는 34개의 속성으로 표현된 사례 집합이다.

엔-버전 학습법이 유전자 학습 환경에 미치는 영향을 평가하기 위해 각각의 사례 집합을 크기가 같은 두 개의 부분 집합인 훈련 사례 집합과 평가 사례 집합으로 나누어 훈련 사례 집합에 대해 엔-버전 학습법을 적용하여 분류 시스템을 구축한 후 평가 사례 집합을 이용하여 구축된 분류 시스템의 성능을 평가하는 과정을 5회 씩 반복하였다. 매회 실험에서는 우선 델보를 이용하여 20개의 분류 규칙 집합을 습득하였다. 분류 규칙 집합의 습득을 위해 사용한 매개 변수의 값은 다음과 같다.

- 개체 집단의 크기 : 15
- 규칙 집합내의 규칙 수의 상한 : 50
- 규칙 집합내의 규칙 수의 하한 : 3
- 교배 연산자 비율 : 98%
- 역치 연산자 비율 : 5%
- 돌연변이 연산자 비율 : 40%

델보를 이용한 분류 규칙 집합의 습득 시 100세대동안 개체 집단의 평균 적합도의 증가량이 0.0005 이하이거나 최대 적합도의 증가량이 0.001 이하이면 학습 과정을 종료하도록 하였으며, 이 경우 분류 규칙 집합의 습득에 소요되는 시간은 평균적으로 붓꽃 사례 집합에 대해서는 400세대, 당뇨 환자의 사례 집합에 대

1) 이들 사례 집합은 "UCI machine learning repository"에 있는 사례 집합들로 ftp를 이용하여 "ics.usi.edu"의 "pub/machine-learning-databases" 디렉토리에서 습득하였음.

해서는 500세대, 그리고 유리 사례 집합과 레이다 시그널 사례 집합에 대해서는 각각 800세대와 1200세대 정도의 시간이 소요되었다. 다음 단계에서는 습득된 20개의 분류 규칙 집합으로부터 3.2절에서 논한 방법에 따라 최적의 분류 규칙 집합의 조합을 구하였다. 최적의 분류 규칙 집합의 조합을 얻기 위해 사용한 매개 변수의 값은 다음과 같으며,

- 개체 집단의 크기 : 100
- 교배 연산자 비율 : 60%
- 돌연변이 연산자 비율 : 30%

세대수가 300세대를 초과하거나 50세대 동안 개체 집단의 최대 적합도에 변화가 없으면 탐색 과정을 종료하도록 하였다.

각각의 사례 집합에 대해 하나의 분류 규칙 집합을 이용하여 분류 시스템을 구축한 경우와 엔-버전 학습법을 이용하여 분류 시스템을 구축한 경우의 학습 결과를 <표 1>에 정리하였다.

〈표 1〉 엔-버전 학습법의 성능 평가 결과

사례집합	훈련사례집합		평가사례집합	
	단일	엔-버전	단일	엔-버전
붓 꽃	98.3	100.0	94.2	96.0
유 리	68.6	80.4	58.0	65.8
당 뇨	78.3	82.0	74.9	76.3
레이다	93.7	98.2	86.6	92.3
평 균	84.7	90.1	78.4	82.6

<표 1>은 단일 분류 규칙 집합을 이용한 분류 시스템이 붓꽃 사례 집합에 대한 실험에서 평균적으로 훈련 사례 집합을 98.3%, 평가 사례 집합내의 사례들의 94.2%를 올바르게 분류한 반면에 엔-버전 학습법 하에서는 훈련 사례 집합을 100.0%, 평가 사례 집합내의 사례들의 96.0%를 올바르게 분류하였음을 보인다. 또한 엔-버전 학습법 하에서 분류 시스템의 성능이 평균적으로 4.2% 향상되었음을 보인다.

다른 학습 알고리즘과의 비교를 위하여 주어진 사례 집합에 대하여 신경망을 이용한 신경 트리를 구축한 후 그 성능을 평가하였다. 실험 결과 붓꽃 사례 집합, 유리 사례 집합, 레이다 시그널 사례 집합에 대해 각각 93.0%, 58.0%, 86.8%의 정확도를 갖는 분류 시스템을 구축할 수 있었다[3]. C4.5 결정 트리 알고리즘을 이용한 경우에는 붓꽃 사례 집합, 유리 사례 집합 및 당뇨 환자 사례 집합에 대해 각각 95.3%, 68.2%, 76.7%의 정확도를 보이는 분류 시스템을 구축할 수 있는 것으로 밝혀졌다[2][15].

5. 결 론

유전자 학습 환경의 학습 능력 향상을 위한 기법으로 엔-버전 학습법을 구현하였다. 엔-버전 학습법은 주어진 사례 집합에 대해 다수의 분류 규칙 집합을 습득한 후 이들 습득된 분류 규칙 집합에 유전자 알고리즘을 적용하여 최적의 분류 규칙 집합의 조합을 습득하고 이를 이용하여 분류 시스템을 구축함으로써 분류 시스템의 성능을 향상시키는 기법이다.

다수의 사례 집합을 이용하여 엔-버전 학습법이 분류 시스템의 성능에 미치는 영향을 평가한 결과 엔-버전 학습법이 분류 시스템의 성능 향상에 크게 기여하는 것으로 나타났다. 본 연구에서는 엔-버전 학습법을 유전자 학습 환경하에서 구현하였으나, 신경망, 결정 트리 등 기존의 학습 알고리즘을 이용한 분류 시스템의 구축 시에도 활용 가능한 기법으로 사료되며, 따라서 분류 시스템의 구축 시 분류 시스템의 성능 향상을 위한 일반적인 기법으로 활용 가능하리라 사료된다.

참 고 문 헌

[1] D. E. Goldberg, *"Genetic Algorithms in Search, Optimization and Machine Learning,"* Addison-Wesley, 1989.

[2] M. Srinivas and L. M. Patnaik, "Genetic Algorithms: A Survey," *IEEE Computer, Vol.27,* pp. 17-26, June, 1994.

[3] C. F. Eick, Y-J. Kim, N. Secomandi, and E. Toto, "DELVAUX An Environment that Learns Bayesian Rule-sets with Genetic Algorithms," *The Third World Congress on Expert Systems,* pp.758-765, Feb., 1996.

2) C4.5의 성능은 주어진 사례 집합의 90%를 훈련 사례로 나머지 10%의 사례를 평가 사례로 하여 평가한 결과 임. 이에 반해 엔-버전 학습법에서는 50%의 사례를 훈련 사례로 나머지 50%의 사례를 평가 사례로 하여 성능을 평가하였음.

[4] C. F. Eick, Y-J. Kim, and N. Secomandi, "Enhancing diversity for a genetic algorithm learning environment for classification tasks," *Int. Conf. Tools with Artificial Intelligence*, pp.820-823, Nov., 1994.

[5] K. De Jong, "Genetic algorithm-based learning," *Machine Learning:An Artificial Intelligence Approach*, Vol.3, Y. Kodratoff and R. S. Michalski, Eds. San Mateo, CA:Morgan Kaufmann, pp.611-638, 1990.

[6] G. Roberts, "Dynamic planning for classifier systems," in *Proc. 5th Int. Conf. Genetic Algorithms*, pp.231-237, 1993.

[7] S. Wilson and D. Goldberg, "A critical review of classifier systems," in *Proc. 3rd Int. Conf. Genetic Algorithms*, pp.244-255, 1989.

[8] J. Oliver, "Discovering individual decision rules: an application of genetic algorithms," in *Proc. 5th Int. Conf. Genetic Algorithms*, pp.216-222, 1993.

[9] A. Homaifar and E. McCormick, "Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms," *IEEE Trans. Fuzzy Systems, Vol.3, No.2*, pp. 129-139, 1995.

[10] C. K. Chiang, H. Y. Chung, and J. J. Lin, "A self-learning fuzzy logic controller using genetic algorithms with reinforcements," *IEEE Trans. Fuzzy Systems, Vol.5*, pp.460-467, 1997.

[11] M. G. Cooper and J. J. Vidal, "Genetic design of fuzzy controllers : the cart and jointed-pole problem," *The Third Int. Conf. Fuzzy Systems, Vol. 3*, pp.1332-1337, 1994.

[12] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, "Acquisition of fuzzy classification knowledge using genetic algorithms," in *Proc. 3rd IEEE Conf. Fuzzy Systems*, pp.1963-1968, 1994.

[13] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka,"Selecting fuzzy if-then rules for classification problems using genetic algorithms," *IEEE Trans. Fuzzy Systems, Vol.3*, pp.250-270, 1995.

[14] H. T. Murata and H. Ishibuchi, "Adjusting membership functions of fuzzy classification rules by genetic algorithms," *Proc. IEEE Int. Conf. Fuzzy Systems*, pp.1819-1824, 1995.

[15] R. Kohavi and G. H. John, "Automatic Parameter Selection by Minimizing Estimated Error," in *Proc. 12 th Int. Conf. Machine Learning*, 1995.

김 영 준

e-mail : yjkim@pine.sangmyung.ac.kr

1984년 고려대학교 산업공학과 졸업(학사)

1990년 미국 Univ. of Houston 전산학(석사)

1996년 미국 Univ. of Houston 전산학(박사)

1997년~현재 상명대학교 정보통신학부 조교수

관심분야 : 기계학습, 유전자 알고리즘, 전문가시스템

홍 철 의

e-mail : hongch@pine.sangmyung.ac.kr

1985년 한양대학교 금속공학과 졸업(학사)

1989년 미국 New Jersey Institute of Technology 전산학(석사)

1992년 미국 Univ. of Missouri-Rolla 전산학(박사)

1997년~현재 상명대학교 정보통신학부 조교수

관심분야 : 기계학습, 최적화 기법, 분산처리