유전 알고리즘 기반 귀납적 학습 환경에서 분류기의 통합

김영준*

Integrating Multiple Classifiers in a GA-based Inductive Learning Environment

Yeongjoon Kim*

이 논문은 2003년도 상명대학교 자연과학연구소 학술연구비 지원으로 수행되었음

요 약

PROSPECTOR에서 사용한 규칙 형태의 분류 규칙을 습득하기 위한 유전 알고리즘 기반 귀납적 학습 환경에서 다중 분류기 학습법을 구현하였다. 다중 분류기 학습법은 주어진 사례 집합에 대해 다수의 분류기를 습득한 후 이를 이용하여 분류 시스템을 구축함으로써 시스템의 성능을 향상시키는 기법이다. 다중 분류기 학습법의 구현을 위해서는 분류기의 분류 결과를 취합하여 최종 결론을 도출해 내기 위한 기법이 필요하다. 본 논문에서는 각각의 클래스에 대해 분류기가 제공하는 사후 가능성을 취합하여 결론을 도출해 내는 기법과 순위에 기반을 둔 보우팅 기법을 소개하고 다중 분류기 학습법이 유전 알고리즘 기반 귀납적 학습 환경에 미치는 영향을 다수의 사례 집합을 이용하여 평가하였다.

ABSTRACT

We have implemented a multiclassifier learning approach in a GA-based inductive learning environment that learns classification rules that are similar to rules used in PROSPECTOR. In the multiclassifier learning approach, a classification system is constructed with several classifiers that are obtained by running a GA-based learning system several times to improve the overall performance of a classification system. To implement the multiclassifier learning approach, we need a decision-making scheme that can draw a decision using multiple classifiers. In this paper, we introduce two decision-making schemes: one is based on combining posterior odds given by classifiers to each class and the other one is a voting scheme based on ranking assigned to each class by classifiers. We also present empirical results that evaluate the effect of the multiclassifier learning approach on the GA-based inductive learning environment.

키워드

다중 전략 학습, 유전 알고리즘, PROSPECTOR, 다중 분류기 학습법

* 상명대학교 소프트웨어학부

접수일자 : 2005. 11. 2

Ⅰ. 서 론

학습 시스템은 시스템의 구현을 위해 이용한 학습 알고리즘의 수에 따라 단일 전략 학습 시스템과 다중 전략 학습 시스템으로 구분된다. 단일 전략 학습 시스템은 결정 트리나, 신경망, 결정 규칙 등의 여러 계산 메커니즘과 귀납법, 연역법, 유추 등의 추론 방법 중에서 하나의 계산 메커니즘과 추론 방법에 기반을 둔 학습 알고리즘을 이용하여 구축된다. 이에 반해 다중 전략 학습 시스템에서는 다수의 추론 형태와 계산 메커니즘을 이용하여 학습 시스템이 구축 된다 [1][2]. 다중 전략 학습 시스템에 대한 연구의 근본적인 목적은 여러 다른 학습 알고리즘이 갖고 있는 고유의 학습 능력을 적절히 통합하여 서로 보완적인 작용을 하게 함으로써 학습 능력이 보다 향상된 시스템을 구축하고자 하는 것이다. 이러한 다중 전략 학습 시스템에 관한 연구의 일환으로 주어진 사례 집합에 대해 서로 다른 학습 알고리즘을 이용하여 사례들을 분류할 수 있는 분류기를 습득한 후 이들 분류기를 통합하여 보다 성능이 향상된 분류 시스템을 구축하기 위한 연구가 최근까지 활발히 진행되어 왔다 [3][4].

유전 알고리즘[5]은 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. 유전 알고리즘은 일반적인 탐색 문제 및 여러 최적화 문제 등의 해결에 널리 이용되어 왔으며 기계 학습 분야에서는 다양한 학습 시스템의 구축과 함께 생성 규칙의 습득[6], 퍼지 컨트롤러와 분류 시스템의 구현을 위한 퍼지 규칙의 습득[7][8] 등에 이용되었다.

유전 알고리즘 기반 학습 환경 하에서는 난수 발생기에 의존하는 탐색과정으로 인해 난수 발생기에 사용하는 초기 값을 달리 함에 따라 학습 시스템은 다른 탐색 공간을 탐색하게 되어 결과적으로 다른 학습 결과를 제공하게 된다. 본 논문에서는 이러한 유전 알고리즘 기반 학습 환경의 특성을 이용하여 주어진 사례의 집합으로부터 다수의 분류기를 습득한 후 이를 이용하여 분류 시스템을 구축함으로써 시스템의 분류 성능을 향상 시키는 다중 분류기 학습법에 관해 연구하였다.

본 논문의 구성은 다음과 같다. 2장에서는 분류 규칙을 습득하기 위한 유전 알고리즘 기반 귀납적 학습 환경을 소개하고 3장에서는 다중 분류기 학습법에 대해 설명한다. 4장에서는 다중 분류기 학습법의 구현에 필요한 의사결정기법을 제시하고 5장에서는 다중 분류기 학습법을 위한 유전 알고리즘 기반 학습 환경에 적합한 적합도 함수를 제시한다. 6장에서는 다중 분류기 학습법이 유전 알고리즘 기반 학습 환경의 학습 능력 향상에 미치는 영향을 평가하고 7장은 결론 및 향후 과제에 대해 설명한다.

Ⅱ. 분류 규칙 습득을 위한 유전 알고리즘 기반 학습 환경

분류 규칙의 습득을 위한 유전 알고리즘 기반 학습 환경에서 훈련 사례 집합 내의 각각의 사례는 속성 $A_1, A_2, \cdots, A_n$에 대한 값 $a_1, a_2, \cdots, a_n$과 사례가 속한 클래스 c로 구성된 리스트, $(a_1, a_2, \cdots, a_n, c)$의 형태로 표현된다. 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. 이러한 정규화 과정은 사례가 규칙의 조건을 만족시키는 정도에 따라 규칙의 결론에 대한 가능성을 증가 혹은 감소시키면서 추론을 수행하는 PROSPECTOR[9]에서 사용한 추론 방법의 적용을 가능하게 한다.

학습 시스템은 주어진 사례의 집합으로부터 "If E then C with S = s, N =n" 형태의 분류 규칙들을 습득한다. 여기서 S와 N은 사례가 규칙의 조건 E를 만족시키는 정도에 따라 규칙의 결론인 클래스 C에 속할 가능성을 증가 혹은 감소시키기 위해 제공되는 가능성에 대한 승수 값의 범위를 나타낸다. 습득된 규칙들로 구축된 분류 시스템에서 각각의 규칙들은 주어진 사례가 규칙의 조건을 완전하게 만족하면(즉, $P(E') = 1$) S의 값을, 불만족 시에는(즉, $P(E') = 0$) N의 값을, $0 < P(E') < 1$인 경우에는 $P(E')$의 값에 비례하여 N과 S사이의 값을 제공한다. 분류 시스템은 각각의 결론 C에 대한 확률 P(C)로부터 사례가 C에 속할 사전 가능성 O(C)를 식 $O(C) = P(C)/(1 - P(C))$에 따라 구하고 이에 C를 결론으로 갖는 규칙들이 제공하는 가능성에 대한 승수를 취합하여 C에 속할 사후 가능성 O(C')을 구한 후 사후 가능성이 가장 큰 클래스를 사례가 속한 클래스로 선택한다. 주어진 사례가 클래스 C에 속할 확률

P(C)는 사례 집합 내에서 C에 속하는 사례가 차지하는 비율로부터 구한다.

분류 시스템이 사례를 분류하는 과정은 다음과 같다:

1. 각각의 클래스 C_k에 대해 사례 집합에서 C_k에 속한 사례의 비율에 따라 $P(C_k)$를 구한 후 사전 가능성 $O(C_k) = P(C_k)/(1 - P(C_k))$를 구한다.
2. C_k를 결론절에서 참조하는 규칙

(r_1) If E_1 then C_k with $S=s_1$, $N=n_1$

...

(r_p) If E_p then C_k with $S=s_p$, $N=n_p$

들이 제공하는 C_k에 대한 승수

$$\lambda_{ri} = \frac{O(C_k|E_i')}{O(C_k)} \qquad for\ i = 1, \ldots, p$$

를 이용하여 사후 가능성

$$O(C_k') = O(C_k|E_1' \wedge \ldots \wedge E_p') = O(C_k) \times \Pi_{i=1}^{p} \lambda_{ri}$$

을 구한다.

3. 사후 가능성이 가장 큰 클래스를 주어진 사례가 속한 클래스로 선택한다.

특정 속성에 대해 사례들이 갖는 값의 상대적인 대소 관계나 속성 값들이 특정 값을 중심으로 하여 분포하는 성향 등은 서로 다른 클래스에 속한 사례들을 분류하는 기준으로 고려해 볼 수 있는 일반적 특징이라 할 수 있다. 이러한 직관적 고찰에 따라 학습 시스템은 주어진 사례 집합으로부터 두 가지 유형의 분류 규칙을 습득한다. 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 사례가 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 C에 대해 N과 S사이의 값을 제공한다. 학습 시스템은 주어진 사례의 집합에 대해 사례들을 분류하기 위해 필요한 속성들을 적절히 고려한 규칙들을 각각의 규칙에 필요한 s, n, 상수 a의 값과 함께 유전 알고리즘을 이용하여 습득하는 것이다.

유전 알고리즘을 이용한 학습 시스템의 구현에서 개체 집단은 일정 수의 규칙 집합으로 구성되며 각각의 규칙 집합은 임의의 수의 분류 규칙으로 구성된다. 초기의 개체 집단은 난수 발생기를 이용하여 임의의 수의 분류 규칙들로 이루어진 일정 수의 규칙 집합을 생성함으로써 얻어진다. 계속되는 진화 과정에서는 적합도에 비례하여 선택된 규칙 집합에 대해 교배 연산자와 돌연변이 연산자 등의 유전 연산자를 적용하여 새로운 개체 집단을 생성하는 과정을 원하는 해가 얻어질 때까지 반복한다. 각각의 규칙 집합의 적합도는 규칙 집합이 사례들을 어느 정도 정확하게 분류 할 수 있는가 하는 분류의 정확도를 이용하여 평가한다.

Ⅲ. 다중 분류기 학습법

2장에서 설명한 학습 환경 하에서 다중 분류기 학습법의 구현은 두 단계로 이루어진다. 첫 단계는 다수의 분류기를 습득하기 위한 단계로 이 단계에서는 주어진 사례 집합에 대해 학습 시스템을 반복 실행하여 다수의 규칙 집합, 즉 분류기를 습득한다. 두 번째 단계에서는 습득한 다수의 분류기에 탐색 알고리즘을 적용하여 최적의 분류기 조합을 찾아낸 후 이를 이용하여 분류 시스템을 구축한다. 그림 1은 유전 알고리즘 기반 학습 환경에서의 다중 분류기 학습법을 보인 것이다.

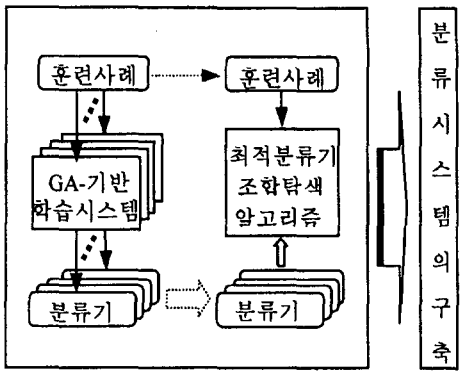

그림 1. 다중 분류기 학습법
Fig. 1 Multiclassifier Learning Approach

본 연구에서는 다중 분류기 학습법의 두 번째 단계를 유전 알고리즘을 이용하여 구현하였다. 유전 알고리즘을 이용한 구현에서 개체 집단의 구성원은 다중 분류기 학습법의 첫 단계에서 습득한 분류기의 수만큼의 이진 문

자로 표현되며 각각의 이진 문자는 상응하는 분류기가 구성원에 포함되는지의 여부에 따라 0과 1로 표현된다. 분류기의 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 분류기의 조합을 습득할 때까지 반복한다. 구성원의 적합도는 구성원 내의 분류기로 구축한 분류 시스템이 훈련 사례 집합에 대해 보이는 분류의 정확도로 평가 한다.

Ⅳ. 다중 분류기를 이용한 의사 결정 기법

다중 분류기 학습법 하에서 구축된 분류 시스템은 다수의 분류기와 의사 결정자로 구성된다 (그림 2 참조). 사례에 대한 분류 과정에서 각각의 분류기는 최종 분류 결과를 도출해 내기 위해 필요한 자료를 의사 결정자에게 제공하고 의사 결정자는 이들 자료를 취합하여 주어진 사례가 속할 클래스를 결정한다.

분류 결과의 취합을 위해 이용 가능한 의사 결정 기법 중 하나는 단순 보우팅을 이용하는 것이다. 이 기법 하에서 분류기는 주어진 사례를 분류한 후 그 결과를 의사 결정자에게 제공하고 의사 결정자는 다수결의 원칙에 따라 사례가 속할 클래스를 최종적으로 선택한다. 이 기법은 구현이 간단한 반면에 사례의 분류 과정에서 분류기가 제공할 수 있는 좀 더 유용한 분류 정보를 의사 결정 과정에 반영

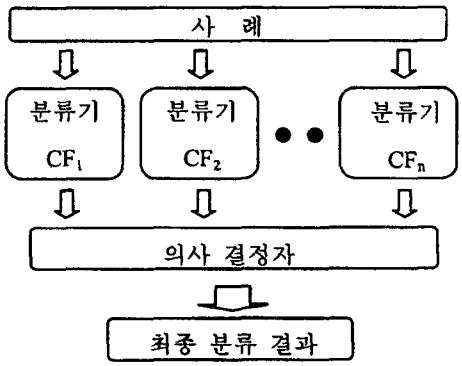

그림 2. 분류 시스템의 구조
Fig. 2 Structure of Classification System

하지 못하는 단점이 있다. 본 논문에서는 단순 보우팅 기법보다는 좀 더 정확한 분류 결과를 얻을 수 있으리라는 기대 하에 사후 가능성의 취합에 기반을 둔 의사 결정 기법과 순위 보우팅에 기반을 둔 의사 결정 기법을 개발하였다.

사후 가능성의 취합에 기반을 둔 의사 결정 기법 하에서 분류기는 주어진 사례에 대해 2장에서 논한 방법으로 구한 사후 가능성들을 식 (1)에 따라 정규화 하여 그 결과를 의사 결정자에게 제공한다.

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

식 (1)에서 $O(C_j')$는 사후 가능성을 나타내고 $\max_k\{O(C_k')\}$는 클래스들에 대한 사후 가능성 중 가장 큰 값을 반환하는 함수를 의미하며 $NO(C_j')$는 클래스 C_j에 대한 정규화된 사후 가능성을 의미한다. 의사 결정자는 정규화 된 사후 가능성들을 각각의 클래스 C_j에 대해 식 (2)에 따라 취합한 후 CE값이 가장 큰 클래스를 사례가 속한 클래스로 선택한다.

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

식 (2)에서 $NO_{CFi}(C_j')$는 분류기 CF_i가 클래스 C_j에 대해 제공하는 정규화된 사후 가능성을, n은 분류 시스템내의 분류기의 수를 나타낸다.

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

순위 보우팅 기반 의사 결정 기법 하에서 분류기는 사례가 클래스에 속할 사후 가능성을 구한 후 사후 가능성이 큰 순서대로 클래스에 순위를 부여하여 이를 의사 결정자에게 제공한다. 의사 결정자는 분류기가 제공하는 순위를 취합하여 그 결과가 가장 적은 것을 사례가 속한 클래스로 결정한다.

V. 다중 분류기 학습법을 위한 적합도 함수

2장에서 설명한 학습 시스템의 구축 시 개체 집단의 구성원인 규칙 집합의 적합도를 평가하기 위한 가장 일반적인 방법은 구성원내의 규칙들로 구축한 분류 시스템이 주어진 사례를 얼마나 정확하게 분류하느냐 하는 분류의 정확도로 평가하는 것이다. 이 경우에 식 (3)과 같은 적합도 함수를 사용할 수 있는데

$$fitness(RS) = |CS| / |TR| \quad (3)$$

식 (3)에서 TR은 훈련 사례 집합을 나타내고 CS는 훈련 사례의 집합에서 규칙 집합 RS에 의해 올바르게 분류된 사례의 집합을 나타낸다.

식 (3)을 이용하여 구현된 유전 알고리즘 기반 학습 환경은 다중 분류기 학습 환경 하에서의 의사 결정 과정에서 고려해야 할 요소들을 분류기의 습득 과정에서 반영하지 않아 다중 분류기 학습 환경에 적합한 분류기를 습득하지 못하는 결과를 초래할 수 있다. 규칙 집합의 적합도 평가에 사후 가능성 취합에 기반을 둔 의사 결정 과정을 반영함으로써 사후 가능성 취합 기반 다중 분류기 학습법에 좀 더 적합한 규칙 집합을 습득하기 위해 적합도 함수를 식 (4)와 같이 개발하였다.

$$fitness'(RS)=(|CS|-\sum_{e\in w}E_{NO}(e)*E_w)/|TR| \quad (4)$$

식 (4)에서 e는 훈련 사례, TR, CS, W는 각각 훈련 사례 집합, 훈련 사례 중 올바르게 분류한 사례의 집합, 잘 못 분류한 사례의 집합을 나타내고 $E_{NO}(e)$은 "$E_{NO}(e) = 1 - $(사례 e가 속한 클래스에 부여한 정규화 된 사후 가능성)"의 식에 따라 계산된 값으로 규칙 집합 RS가 주어진 사례 e를 올바르게 분류한 경우에는 0을, 잘 못 분류한 경우에는 e가 속한 클래스에 대하여 부여해야 할 정규화 된 사후 가능성 1에서 실제로 계산된 사후 가능성을 뺀 결과를 나타낸다. E_w는 잘못된 정도를 나타내는 $E_{NO}(e)$의 값에 비례하여 적합도를 감소시켜주기 위해 정한 임의의 상수이다. 식 (4)에서는 규칙 집합이 사례가 속한 클래스에 대해 부여해야 할 올바른 정규화 된 사후 가능성인 1이 아닌 다른 잘못된 값을 제공하는 경우에 그 차이에 비례하여 적합도가 감소하도록 적합도 함수를 수정하였다.

순위 보우팅 기반 다중 분류기 학습법을 위한 적합도 함수로는 식 (5)를 개발하였다.

$$fitness''(RS)=(|CS|-\sum_{e\in w}E_R(e)*E_w)/|TR| \quad (5)$$

식 (5)에서 $E_R(e)$는 사례 e가 속한 클래스에 부여한 순위에서 1을 뺀 결과로 규칙 집합 RS가 e를 올바르게 분류한 경우에는 0을, 잘 못 분류한 경우에는 e가 속한 클래스에 대하여 부여한 순위에서 실제 부여해야 할 순위를 뺀 결과를 나타낸다. E_w는 잘못된 순위를 부여한 경우에 그에 비례하여 적합도를 감소시키기 위해 정한 임의의 상수이다. 식 (5)에서는 규칙 집합이 사례가 속한 클래스에 대해 부여해야 할 순위인 1이 아닌 다른 잘못된 순위를 제공하는 경우에 순위의 차에 비례하여 적합도가 감소하도록 적합도 함수를 수정하였다.

VI. 다중 분류기 학습법의 성능 평가

다중 분류기 학습법이 분류시스템의 성능 향상에 미치는 영향을 다음의 사례 집합을 이용하여 평가 하였다. (이들 사례 집합은 "UCI machine learning repository" 에서 습득하였음)

- 붓꽃 사례 집합: 3가지 종류의 붓꽃으로부터 얻어진 150개의 사례가 꽃잎의 길이와 넓이, 꽃받침의 길이와 넓이의 4가지 속성 값과 붓꽃의 종류를 나타내는 값으로 표현된 사례 집합
- 유리 사례 집합: 6 종류의 유리 파편들로부터 얻어진 214개의 사례가 유리를 구성하는 9가지 물질의 성분비와 유리의 종류를 나타내는 값으로 표현 됨
- 레이더 시그널 사례 집합: 올바른 경우와 잘못된 경우의 351개 레이더 시그널 사례가 34가지의 속성 값과 사례가 속한 클래스로 표현 된 사례 집합
- 콩의 질병 사례 집합: 콩에 감염될 수 있는 15가지 질병으로부터 얻어진 290개의 사례가 35개의 속성 값과 사례가 속한 클래스로 표현됨
- 당뇨 환자 사례 집합: 당뇨 환자인 경우와 정상인인 경우로 분류되는 768개의 사례가 8가지 속성 값과 사례가 속한 클래스로 표현된 사례 집합

다중 분류기 학습법의 성능을 평가하기 위해 우선 각

각의 사례 집합을 크기가 같은 두 개의 부분 집합인 훈련 사례 집합과 평가 사례 집합으로 나눈 후 훈련 사례 집합에 대해 학습 알고리즘을 반복 실행하여 20개의 분류기를 구한다. 그런 다음 습득된 20개의 분류기의 집합으로부터 유전 알고리즘을 이용하여 최적의 분류기 조합을 찾아내어 이를 이용하여 분류 시스템을 구축 하고 평가 사례를 이용하여 분류 시스템의 성능을 평가한다. 이와 같은 실험을 각각의 사례 집합에 대해 5회씩 반복하였다. 단일 분류기의 학습 과정에서 유전 알고리즘을 위한 적합도 함수는 5장에서 식 (3)으로 주어진 적합도 함수 fitness를 사용하였다.

단일 분류기를 이용하여 구축한 분류 시스템과 다수의 분류기를 이용하여 구축한 분류 시스템의 성능을 표 1에 비교하였다. 표 1에서 '단일'은 하나의 분류기를 이용한 분류 시스템을, '다중'은 다수의 분류기를 이용한 분류 시스템을 나타낸다. '다중'에서 세 개의 열 SV, CO, VR는 각각 의사 결정 기법으로 단순 보우팅, 사후 가능성의 취합, 순위 보우팅 방식에 따른 성능을 평가한 결과를 나타낸다.

표 1. 다중 분류기 학습법의 성능
Table. 1 Performance of Multiclassifier Learning Approach

(a) 훈련 사례 집합
(a) Training Data Set

사례집합	단일	다중		
		SV	CO	VR
붓꽃	98.7	100.0	100.0	100.0
유리	68.6	81.5	81.1	79.2
당뇨	78.0	82.2	81.3	82.2
레이더	92.6	98.2	98.1	98.2
콩의질병	56.6	76.6	76.1	68.4
평 균	78.9	87.7	87.3	85.6

(b) 평가 사례 집합
(b) Testing Data Set

사례집합	단일	다중		
		SV	CO	VR
붓꽃	94.1	95.8	95.1	95.5
유리	57.4	64.3	62.3	62.2
당뇨	74.6	75.2	75.5	75.2
레이더	86.1	91.2	90.7	91.2
콩의질병	50.8	67.8	68.2	60.0
평 균	72.6	78.9	78.4	76.8

표 1은 단일 분류기를 이용한 분류 시스템이 붓꽃 사례 집합에 대한 실험에서 평균적으로 훈련 사례를 98.7%, 평가 사례를 94.1% 올바르게 분류한 반면에 단순 보우팅을 이용한 다중 분류기 시스템은 훈련 사례를100.0%, 평가 사례를 95.8% 올바르게 분류하였음을 보인다. 또한 단순 보우팅에 기반을 둔 다중 분류기의 경우에 평가 사례 집합에 대한 분류 시스템의 성능이 단일 분류기에 비해 평균적으로 6.3% 향상되었음을 보인다.

표 1의 결과에서 보면 다중 분류기 시스템의 구축 시 단순 보우팅이 다른 의사 결정 기법에 비해 좀 더 나은 성능을 보인 것으로 나타났으나 이는 단일 분류기의 습득 시 적합도 함수 fitness를 이용한 것에 기인한다. 적절한 적합도 함수의 사용이 다중 분류기 학습법의 성능에 미치는 영향을 평가하기 위해 적합도 함수 fitness'와 fitness"를 사용하여 앞에서 언급한 방법과 동일한 방법으로 다중 분류기 시스템을 구축하고 성능을 평가하였다.

표 2는 적합도 함수 fitness'를 이용하여 구축한 사후 가능성 취합 기반 다중 분류기 시스템의 성능을 적합도 함수 fitness를 이용한 경우와 비교하여 보인 것이다.

표 2. 성능 비교 : fitness vs. fitness'
Table. 2 Performance Comparison : fitness vs. fitness'

사례집합	fitness		fitness'	
	훈련	평가	훈련	평가
붓꽃	100.0	95.1	100.0	96.0
유리	81.1	62.3	80.4	65.8
당뇨	81.3	75.5	82.0	76.3
레이더	98.1	90.7	98.2	92.3
평 균	90.1	80.9	90.2	82.6

표 2는 적합도 함수 fitness'를 이용하여 구축한 분류 시스템이 유리 사례 집합에 대해 훈련 사례를 80.4%, 평가 사례를 65.8% 올바르게 분류하여 fitness를 이용하여 구축한 분류 시스템에 비해 약 3.5% 분류의 정확도를 향상 시키고 있음을 보인다. 또한 fitness'의 사용 시 사례 집합에 대한 평균 분류 성능은 82.6%로 이는 적절한 적합도 함수의 사용이 다중 분류기 학습법의 성능을 평균적으로 1.7% 더 향상 시키고 있음을 보인다.

표 3은 적합도 함수 fitness"를 이용하여 구현한 순위 보우팅 기반 다중분류기 시스템의 성능을 fitness를 이용한 경우와 비교하여 보인 것이다. 순위 보우팅 기법은 사례가 속할 수 있는 클래스의 개수가 다수인 경우에 효과적

인 기법이므로 붓꽃, 당뇨, 레이더의 사례 집합과 같이 사례가 속할 클래스의 수가 적은 사례 집합은 평가에서 제외하였다.

표 3. 성능 비교 : fitness vs. fitness"
Table. 3 Performance Comparison : fitness vs. fitness"

사례집합	fitness		fitness"	
	훈련	평가	훈련	평가
유 리	79.2	62.2	79.0	66.3
콩의질병	68.4	60.0	75.9	68.4
평 균	73.8	61.1	77.5	67.4

표 3은 적합도 함수 fitness"를 이용하여 구축한 분류 시스템이 유리 사례 집합에 대해 훈련사례를 79.0%, 평가사례를 66.3% 올바르게 분류하여 fitness를 이용하여 구축된 분류 시스템에 비해 약 4.1% 분류의 정확도를 향상시키고 있음을 보인다. 또한 두 개의 사례 집합에 대해 평균적으로는 6.3% 정확도를 향상시킴을 보인다.

본 연구에서 구축한 분류 시스템의 성능을 신경망과 결정 트리 알고리즘을 이용하여 구축된 분류 시스템의 성능과 비교하였다. 표 4에서 '다중'은 본 논문에서 제시한 다수의 분류기를 이용하여 구축한 분류 시스템의 성능을 나타내고 '신경망'과 'C4.5'는 각각 신경망과 C4.5 결정 트리 알고리즘을 이용하여 다수의 분류기를 구한 후 이를 이용하여 구축한 분류 시스템의 성능을 보인 것이다.

표 4. 학습 시스템의 성능 비교
Table. 4 Performance Comparison with Other Learning Systems

사례 집합	다중	신경망	C4.5
붓꽃	96.0	95.9	95.9
당 뇨	76.3	76.8	72.4

표 4의 결과는 본 연구를 통해 구축한 학습 시스템이 신경망과 결정 트리 알고리즘에 견줄만한 학습 능력이 있음을 보인다.

Ⅶ. 결 론

유전 알고리즘 기반 학습 환경의 학습 능력 향상을 위한 기법으로 다중 분류기 학습법을 구현하였다. 다수의 사례 집합을 이용하여 분류 시스템의 성능에 미치는 영향을 평가한 결과 다중 분류기 학습법이 분류 시스템의 성능 향상에 크게 기여하는 것으로 나타났다. 각각의 의사 결정 기법에서 고려하는 요소들을 분류기의 습득 과정에서 반영할 수 있도록 적절한 적합도 함수를 개발하여 구현 한 결과는 적합한 학습 환경의 구현이 다중 분류기 학습법의 성능을 더욱 더 향상시킴을 보인다.

다중 분류기 학습법의 구현을 위해서는 분류 시스템 내의 분류기가 적정 수준의 다양성을 유지할 필요가 있다. 여기서 분류기의 다양성이란 분류기의 귀납적 기울어짐(inductive bias)의 다양성을 의미하며 이는 의사 결정 과정에서 주어진 사례에 대해 타 분류기와는 다른 속성들을 고려하여 결론을 도출해 내는 요인이 된다. 따라서 이러한 귀납적 기울어짐의 다양성을 적절히 유지하도록 하여 분류 시스템을 구축하면 시스템의 성능을 좀 더 향상시킬 수 있을 것이다. 향후에는 분류기의 다양성을 평가하기 위한 기법의 개발과 함께 분류기의 습득 및 최적 분류기 조합의 습득 과정에서 분류기의 다양성을 반영하기 위한 연구가 이루어져야 하겠다.

참고문헌

[1] F. Esposito, G. Semeraro, N. Fanizzi, and S. Ferilli, "Multistrategy Theory Revision: Induction and Abduction in INTHELEX," *Machine Learning*, 38, pp. 133-156, 2000.

[2] C. Giraud-Carrier, R. Vilalta, and P. Brazdil, "Introduction to the Special Issue on Meta-Learning," *Machine Learning*, 54(3), pp. 187-193, 2004.

[3] G. Rätsch, M. K. Warmuth,, "Maximizing the margin with boosting," *Annual Conference on Computational Learning Theory*, LNAI 2375, pp. 334-350, 2002.

[4] R. L. Major and C. T. Ragsdale, "An aggregation approach to the classification problem using multiple prediction experts," *Information Processing and Management*, 36, pp. 683-696, 2000.

[5] M. Srinivas and L. M. Parnaik, "Genetic algorithms: a survey," *IEEE Computer*, Vol. 27, pp. 17-26, June 1994.

[6] G. Roberts, "Dynamic planning for classifier systems," in

Proc. 5th Int. Conf. Genetic Algorithms,pp. 231-237, 1993.

[7] 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.

[8] H. Ishibuchi and T. Nakashima, "Improving the Performance of Fuzzy Classifier Systems for Pattern Classification Problems with Continuous Attributes," *IEEE Transactions on industrial electronics*, Vol. 46, No. 6, pp. 1057 – 1068, December 1999.

[9] R. Duda, P. Hart and J. Nilsson, "Subjective Bayesian methods for rule-based inference systems,' in *Proc. National Computer Conference*, pp. 1075 – 1082, 1976

저자소개

김 영 준 (Yeongjoon Kim)

1984년 고려대학교 산업공학과 (공학사)
1996년 미국 Univ. of Houston 전자계산학과(박사)
1997년~현재 상명대학교 소프트웨어학부 부교수
※관심분야 : 기계학습, 진화알고리즘, 전문가시스템