분류 성능 향상을 위한 다양성 기반 앙상블 유전자 프로그래밍

(Diversity based Ensemble Genetic Programming for Improving Classification Performance)

홍 진 혁 † 조 성 배 ††
(Jin-Hyuk Hong) (Sung-Bae Cho)

요 약 분류 성능을 향상시키기 위해서 다수의 분류기들을 결합하는 연구가 활발히 진행되고 있다. 우수한 앙상블 분류기를 획득하기 위해서는 정확하고 다양한 개별 분류기를 구축해야 한다. 기존에는 Bagging이나 Boosting 등의 앙상블 학습 기법을 이용하거나 획득된 개별 분류기의 학습 데이타에 대한 다양성을 측정하였지만 유전 발현 데이타와 같이 학습 데이타가 적은 경우 한계가 있다. 본 논문에서는 유전자 프로그래밍으로부터 획득된 규칙의 구조적 다양성을 분석하여 결합하는 앙상블 기법을 제안한다. 유전자 프로그래밍으로 해석 가능한 분류 규칙을 생성하고 그들 사이의 다양성을 측정한 뒤, 이들 중 다양한 규칙의 집합을 결합하여 분류를 수행한다. 유전 발현 데이타로부터 림프종 암, 폐 암, 난소 암 등을 분류하는 문제를 대상으로 실험하여 제안하는 방법의 유용성을 검증하였다. 앙상블 시 분류 규칙 사이의 다양성을 분석하여 결합한 결과, 다양성을 고려하지 않을 때보다 높은 분류 성능을 획득하였고, 개별 분류 규칙들 사이의 다양성에 따라서 정분류율이 증가하는 것도 확인하였다.

키워드 : 유전자 프로그래밍, 앙상블, 다양성, 암 분류

Abstract Combining multiple classifiers has been actively exploited to improve classification performance. It is required to construct a pool of accurate and diverse base classifier for obtaining a good ensemble classifier. Conventionally ensemble learning techniques such as bagging and boosting have been used and the diversity of base classifiers for the training set has been estimated, but there are some limitations in classifying gene expression profiles since only a few training samples are available. This paper proposes an ensemble technique that analyzes the diversity of classification rules obtained by genetic programming. Genetic programming generates interpretable rules, and a sample is classified by combining the most diverse set of rules. We have applied the proposed method to cancer classification with gene expression profiles. Experiments on lymphoma cancer dataset, prostate cancer dataset and ovarian cancer dataset have illustrated the usefulness of the proposed method. A higher classification accuracy has been obtained with the proposed method than without considering diversity. It has been also confirmed that the diversity increases classification performance.

Key words : genetic programming, ensemble, diversity, cancer classification

1. 서 론

암 분류는 정확한 치료와 진단을 위해 매우 중요하다. 임상에 의존적인 기존의 방법은 다양한 종의 암을 정확히 분간하는데 한계가 있는데, 최근 유전 발현 데이타를 이용한 질병 및 암의 진단이 주목받고 있다[1]. DNA 마이크로어레이 기술을 이용하여 수집하는 유전 발현 데이타는 인체 기관이나 질병 등에 관련된 다양한 정보를 제공하여 보다 효과적인 암 진단에 도움을 주고 있다[2]. 일반적으로 유전 발현 데이타는 매우 많은 양의 정보를 포함하기 때문에, 이를 직접 분석하는 것은 거의 불가능하다. 따라서 신경망[3], 베이지안 방법[4], SVM [5], 결정트리[6] 및 k 최근접 이웃[7] 등의 많은 기계학습 기법이 활발히 적용되고 있다.

진화 기술도 유전 발현 데이타를 분석하는데 많이 사

· 이 연구는 산업자원부가 지원한 뇌과학 연구 프로그램에 의해 지원되었음
† 학생회원 : 연세대학교 컴퓨터과학과
hjinh@sclab.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@sclab.yonsei.ac.kr
논문접수 : 2005년 3월 10일
심사완료 : 2005년 10월 2일

용되고 있다. Li 등은 효과적인 유전자 선택을 위해 유전자 알고리즘과 k 최근접 이웃의 결합 모델을 제안하였고[7], Deutsch는 최적의 유전자 집합을 찾기 위해 진화 알고리즘을 사용하였다[8]. Karzynsci 등은 신경망과 유전자 알고리즘을 결합하여 암을 진단하였고[9], Langdon과 Buxton은 DNA 칩 데이타를 분류하는데 유전자 프로그래밍을 적용하였다[10]. 뿐만 아니라 높은 암 분류율을 얻기 위해 다수의 분류기를 결합하는 앙상블 기법도 Valentini[11], Park과 Cho[12], Tan과 Gilbert[13] 등에 의해 시도되었다.

유전 발현 데이타는 보통 고차원의 특징을 가지는 적은 수의 샘플로 구성되어 있기 때문에 많은 기계학습 기법들이 과잉 학습(overfit)되기 쉽다. 앙상블은 그러한 문제에 대해서 분류 성능과 함께 안정성을 높여준다. 앙상블 기법은 다수의 분류기를 효과적으로 결합하여 개별 분류기보다 정확하고 안정적인 분류 성능을 얻는 것으로, 다양한 개별 분류기를 생성하고 이들 분류기를 효과적으로 결합하는 방법에 대해서 활발히 연구되고 있다[14,15].

앙상블에 있어서 다양한 개별 분류기를 확보하는 것은 매우 중요하다. 이미 알려져 있듯이, 동일한 분류기의 결합으로는 성능향상이 불가능하며, 개별 분류기들이 서로 다를 때에 성능이 향상될 수 있다. 이상적으로는 개별 분류기의 분류율이 50% 이상만 된다면 개별 분류기의 수가 늘어날수록 분류율이 높아져 100%에 도달한다고 증명되었으나, 실제로는 다양성과 개별 분류기의 분류율 사이에 이율 배반성이 존재하기 때문에 성능이 계속해서 향상되지는 않는다[16]. 많은 연구자들은 다양한 개별 분류기를 생성하여 이런 한계를 극복하고자 하는데, 대표적인 앙상블 학습 기법으로 Bagging과 Boosting이 있다. Breimen에 의해 소개된 Bagging(Bootstrap aggregating)은 원 데이타로부터 임의로 구성된 학습 데이타를 사용하여 개별 분류기를 학습하며, Schapire에 의해 제안된 Boosting은 일련의 개별 분류기를 학습한다. 이전 분류기에서 잘못 분류된 샘플은 다음 분류기의 학습 데이타에 높은 확률로 참여한다. Arching이나 Ada-Boosting은 최근 많이 사용되는 Boosting 기법이다[16,17].

다양한 개별 분류기를 생성하기 위한 많은 연구가 수행되었는데, Webb과 Zheng은 다전략 앙상블 학습 기법을 제안하였고[16], Optiz와 Maclin은 대표적인 앙상블 학습 기법에 대한 실험적 연구를 수행하였다[17]. Bryll은 임의의 특징 집합으로 분류기를 학습하는 변수 Bagging을 제안하였고[15], Islam은 서로 음의 속성을 가지는 신경망 집합을 학습하여 결합하였다[18]. 그 외에도 많은 연구가 보다 다양하고 정확한 개별 분류기 생성을 위해 진행되었으며, 특히 개별 분류기 사이의 다양성을 측정하여 결합 시 활용하는 기법도 연구되었다[19].

앙상블에서 중요한 또 다른 문제는 개별 분류기를 어떻게 결합하는가이다. 비록 우수한 개별 분류기를 획득하였더라도 적절한 결합 전략을 채택하지 않으면 높은 앙상블 효과를 기대할 수 없다. 단순한 결합 전략으로는 다수결 투표, 평균치, 가중치 평균값, 최소값, 최대값 및 Borda 계수 등이 있으며, 이들은 입력된 샘플에 대한 분류기의 결과만을 고려하여 분류를 수행한다. 이와는 달리, Naive Bayes, 행위지식 공간(Behavior knowledge space), 결정 템플릿(decision template), Dempster-Shafer 결합 및 퍼지 적분 등의 보다 정교한 결합 방법은 학습 과정에서 먼저 분류를 위한 모형을 구축한 후 분류를 수행한다. 하나 이상의 개별 분류기가 분류를 제대로 수행하였을 때, 분류를 수행한 것으로 간주하는 Oracle 전략은 분류의 상향선을 설정하는 데 많이 사용된다.

본 논문에서는 이전 연구[20]에서 제안한 유전자 프로그래밍기반 분류규칙을 효과적으로 결합하기 위해 앙상블에서 매우 중요한 개별 분류기의 다양성을 분석하고 유전자 프로그래밍에서의 다양성을 고려한 효과적인 앙상블 전략을 제안한다. 유전자 프로그래밍을 이용하여 개별 분류 규칙을 생성하고, 그들 중 가장 다양한 규칙들을 선택하여 앙상블 분류기로 결합한다. 출력 단계에서 다양성을 측정하던 기존의 방법과는 달리, 유전자 프로그래밍으로부터 획득된 해석가능한 분류 규칙을 직접 비교분석하여 다양성을 측정한다. 대표적인 유전 발현 데이타인 림프종 암 데이타, 난소 암 데이타와 폐 암 데이타로 제안하는 방법의 유용성을 평가하였다.

2. 암 분류를 위한 다양성 기반 앙상블

Koza가 제안한 유전자 프로그래밍은 주어진 문제를 푸는 프로그램을 자동으로 생성한다[21]. 많은 면에서 유전자 알고리즘과 유사하지만 트리 구조를 개체의 표현형으로 사용한다는 점에서 다르다. 하나의 개체는 함수와 말단 변수로 구성된 트리로 표현되며, 적용 문제에 따라서 다양한 변수와 함수가 사용될 수 있다. 유전자 프로그래밍을 이용한 앙상블에 관한 연구가 수행되었는데, Zhang과 Bhattacharyya는 US Air Force Lan 데이타를 분류하기 위해 유전자 프로그래밍을 사용하였고[22], Brameier와 Banzhaf는 유전자 프로그래밍을 이용하여 분류기 집합을 생성하고 몇몇 결합 전략을 사용하여 이들을 결합하였다[23]. Fernandez 등은 실험적으로 다종의 유전자 프로그래밍 연구를 진행하였고[24], Imamura 등은 앙상블 유전자 프로그래밍에서 행위적 다양성을 적용하기도 하였다[25].

제안하는 방법은 그림 1과 같이 크게 두 단계로 구성된다. 유전자 프로그래밍을 이용하여 다수의 개별 분류 규칙을 생성한 후, 이들 규칙 사이의 다양성을 고려하여 결합한다. 먼저 특징 선택 과정으로 데이타의 차원을 축소한 후 유전자 프로그래밍으로 개별 분류 규칙을 생성한다. 다수의 개별 분류 규칙을 반복해서 생성하고, 이들 사이의 다양성을 측정하여 다양성이 최대가 되는 분류 규칙을 선별해 결합한다. 다양성을 측정하기 위한 기존의 방법은 학습 데이타에 대한 개별 분류기들의 출력값을 이용하기 때문에, 측정되는 다양성은 학습 데이타의 속성에 의존적일 수밖에 없다. 이와는 달리, 본 논문에서는 분류 규칙의 구조와 내용을 직접 비교하여 다양성을 측정한다. 따라서 개별 분류기의 결정면이 왜곡되지 않으며, 학습 데이타로부터 발생하는 부작용을 고려할 필요가 없다. 뿐만 아니라 정확한 다양성 측정을 위해 다수의 학습 샘플이 필수적인 기존의 방법과는 달리 학습 데이타를 이용하지 않고 다양성을 측정할 수 있다.

2.1 다중 분류 규칙 생성

DNA 마이크로어레이 데이타는 수천에서 수만의 유전 발현 정보를 포함하고 있으나, 특정 암과 관련된 것은 일부이다. 유용한 유전자 집합은 특징 선택을 통해 구성되며, 특징의 수를 적절한 수로 줄이는 것은 분류 성능을 향상시키는데 필수적이다.

본 논문에서는 암 분류와 관련된 유전자를 선택하기 위해 그림 2에서와 같이 두 개의 대표마커를 정의한다. 이들 대표마커는 서로 음의 관계를 가지고 다른 결정 경계를 표현하며, 특징 선택으로 대표마커와의 유사성을 분석하여 가장 유사한 유전자가 선택된다. 첫 번째 대표

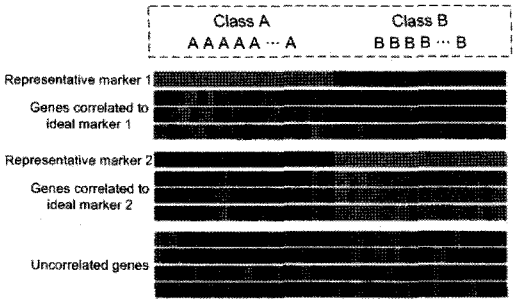

그림 2 음의 상관성을 가지는 유전자 선택

마커는 부류 A에 대해서는 높은 값을, 부류 B에 대해서는 낮은 값을 가지며, 두 번째 대표마커는 부류 A에 대해서는 낮은 값을, 부류 B에 대해서는 높은 값을 갖는다. 전자는 부류 A의 샘플에 대해서는 1로, 부류 B의 샘플에 대해서는 0으로 구성된 이진벡터이며, 후자는 부류 A의 샘플에 대해서는 0으로, 부류 B의 샘플에 대해서는 1로 구성된 이진벡터이다. 모두 길이는 샘플의 수와 같다. 대표마커와의 유사성을 측정하기 위해서 유전자 선택 시 많이 사용되는 코사인 계수를 사용하며, 각 대표마커에 대해 가장 유사한 25개의 유전자를 뽑아 총 50개의 유전자를 선택하였다. 코사인 계수는 다음의 수식과 같이 구한다.

$$CC = \frac{\sum_{i=1}^{n} ideal_i \times g_i}{\sqrt{\sum_{i=1}^{n} ideal_i^{\,2} \times \sum_{i=1}^{n} g_i^{\,2}}}, \quad (1)$$

단, n은 샘플 수

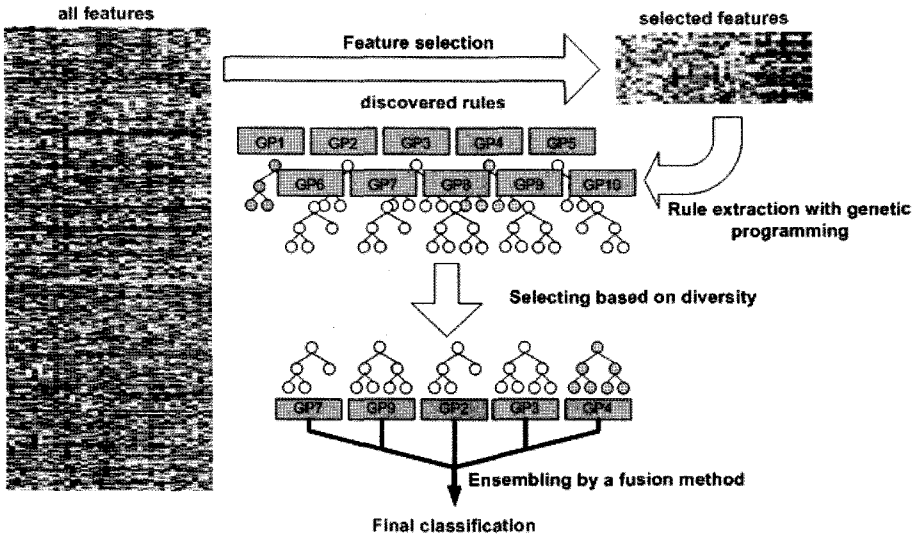

그림 1 제안하는 방법

선택된 유전 발현 정보로 구성된 데이타를 바탕으로 유전자 프로그래밍을 이용해 암 분류 규칙을 생성하는데, 이때 유전자 발현을 고려하여 산술 연산자를 이용한다. 유전자 프로그래밍은 임의로 개체들을 초기화한 후, 평가 단계에서 주어진 문제에 대한 개체의 적합도를 적합도 함수를 통해 계산한다. 다음 세대를 구성할 집단을 생성하기 위해 적합도에 따라서 개체를 선택하며 적자생존의 원칙에 따라 보다 나은 개체가 다음 세대를 위해 선택될 확률이 높도록 한다. 선택된 개체들에 교차나 돌연변이 등의 유전자 연산을 적용하여 다음 세대를 구성하는 새로운 개체들을 생성한다. 유전자 프로그래밍은 평가, 선택과 유전자 연산 과정을 원하는 개체가 획득될 때까지 반복한다.

유전자 프로그래밍에서 개체들은 함수 {+,-,×,÷}와 말단 변수 {$f_1 f_2,...f_n$, 상수}로 구성된 트리로 표현된다. 본 논문에서 사용된 산술 연산자들은 상향적 유전 발현과 하향적 유전 발현을 표현하도록 한다. 분류 규칙은 G={V={EXP, OP, VAR}, T={+,-,×,÷, f_1, f_2, ..., f_n, 상수}, P, {EXP}}이며, 규칙 집합 P는 다음과 같다.

EXP→EXP OP EXP | VAR

OP→+|-|×|÷

VAR→ $f_1|f_2|...|f_n|$상수

분류 규칙을 이용해 샘플을 평가하며 0보다 큰 값이 나오면 부류 A로, 0보다 작은 값이 나오면 부류 B로 분류를 수행한다. 유전자 프로그래밍의 대표적 유전자 연산인 교차, 돌연변이와 치환을 사용한다. 교차는 두 개체의 하위 트리를 임의로 선택하여 교환하는 것이고, 돌연변이는 한 개체의 하위 트리를 임의의 트리로 바꾸는 것이다. 치환은 한 개체 내의 두 개의 하위 트리를 교환하는 것으로 유전 형질은 유지하고 위상 정보만 바꾸는 것이다. 모든 유전자 연산은 미리 설정된 확률에 따라 작동된다.

개체의 평가 과정에서 분류 성능과 규칙의 단순성을 고려한다. 규칙의 단순성은 Occam's razor의 개념으로 설명할 수 있다[23]. 정분류율은 학습 데이타에 대해서 얼마나 정확히 분류를 수행했는지를 측정하고, 단순성은 규칙에 사용된 노드의 수를 의미한다. 본 논문에서는 개체의 유용성을 나타내는 적합도를 다음의 식과 같이 구하며, 정분류율과 단순성에 대한 가중치를 각각 0.9와 0.1로 설정하였다.

$$\text{적합도} = \frac{\text{정확히 분류한 샘플수}}{\text{전체 데이타 샘플수}} \times \text{가중치}_1 + \text{규칙의 단순성} \times \text{가중치}$$

$$\text{규칙의 단순성} = \frac{\text{규칙의 노드수}}{\text{최대 노드수}} \times \text{가중치}_2 \qquad (2)$$

2.2 다양한 분류 규칙 결합

앞서 언급하였듯이, 앙상블의 성능을 향상시키기 위해서는 개별 분류기들이 다양해야한다. Q-statistic, 상관계수, Kohavi-Wolpert 변수 및 엔트로피 등의 다양한 다양성 측정 방법이 있으며, 많은 연구자들이 보다 정확한 다양성 측정을 위한 방법을 연구하고 있다. Zenobi와 Cunningham은 다른 특징 집합에서 발생하는 다양성을 이용하였고[26], Shipp과 Kuncheva는 결합 기법과 다양성 측정 방법 사이의 관계를 분석하였다[19]. Kuncheva와 Whitaker는 앙상블에서의 다양성 측정 방법의 성능을 비교하기도 하였고[27], Windeatt는 다중 분류 시스템을 위한 다양성 측정 방법에 대한 실험적 분석을 시도하였다[28].

앙상블을 위한 기존의 다양성 측정 방법은 정확성-다양성 딜레마(accuracy-diversity dilemma)라고 알려진 정확성과 다양성 사이의 이율 배반성으로 인해 개별 분류기의 정확성이 높아질수록 이들 사이의 다양성은 떨어진다. 만약 개별 분류기들이 모든 학습 샘플을 정확히 분류할 경우에는 다양성이 0이 되어버려 앙상블의 효과가 사라진다. 뿐만 아니라 유전 발현 데이타와 같이 샘플이 적은 데이타는 기존의 방법으로 다양성을 정확히 측정하기 어려운 문제가 있다.

제안하는 방법은 기존의 다양성 측정 방법과는 달리 개별 분류기의 구조를 직접 비교하여 분류 규칙 사이의 다양성을 측정한다. 따라서 다양성 측정 단계에서는 학습 샘플이 필요 없으며, 정확성-다양성 딜레마에 의한 부작용이 발생하지 않는다. 개별 분류기 사이의 구조를 직접 비교하기 위해서는 분류기의 해석이 가능해야 한다. 유전자 프로그래밍은 이를 위해 트리 구조의 해석 가능한 분류 규칙을 생성한다. 의사-동형(pseudo-isomorphs)이나 조작 거리(edit distance) 등의 방법이 트리의 구조를 비교하기 위해 사용되고 있다[29]. 본 논문에서는 단순화된 조작 거리를 이용하여 분류 규칙 r_i와 r_j 사이의 차이를 다음과 같이 측정한다.

$$\text{조작거리}(r_i, r_j) = \begin{cases} d(r_i, r_j), & r_i\text{와 } r_j\text{가 자식이 없을 때} \\ d(r_i, r_j) & + \text{조작거리}(r_i\text{의 RS}, r_j\text{의 RS}) \\ & + \text{조작거리}(r_i\text{의 LS}, r_j\text{의 LS}) \\ & \text{그렇지 않을 때 (단, RS는 오른쪽 자식, LS는 왼쪽 자식)} \end{cases} \qquad (3)$$

$$\text{단, } d(p,q) = \begin{cases} 1, & p = q \\ 0, & p \neq q \end{cases}$$

분류 규칙에 사용되는 유전자도 다양성 측정에 이용된다. 두 분류 규칙이 동일한 유전자를 사용할 경우에는 다양성이 떨어지고, 다른 유전자를 사용할 경우 다양성이 증가한다. 본 논문에서는 10개의 분류 규칙에서 5개의 가장 다양한 분류 규칙을 그림 3의 과정을 통해 선택하여 결합한다. 결합 전략은 다수결 투표(MAJ)[14]와

결정 템플릿(DT)[30]을 이용하였다. 다수결 투표는 각 분류기의 결과 중 가장 많이 선택된 부류를 선택하는 방법이고, 결정 템플릿은 각 클래스에 대한 분류기 출력의 프로파일을 바탕으로 유사도를 비교하여 분류를 수행하는 방법으로 분류기 결합에 대표적인 방법들이다.

```
R: A set of extracted rules {r1, r2, ..., rm}
S: A set of selected rules {s1, s2, ..., sn}

int calculate_diversity(ri, rj) {
  cfij = common_feature_number(ri, rj);
  dfij = different_feature_number(ri, rj);
  edij = edit_distance(ri, rj);
  return dfij - cfij - α × edij
}
For i=1 to m
  For j=i+1 to m
    dij = calculate_diversity(ri, rj);
Find a set S in which rules' diversity is maximized
S = {s1, s2, ..., sn}
```

그림 3 분류 규칙 선택 알고리즘

2.3 다양성 기반 앙상블의 이론적 고찰

학습 샘플과 트리를 구성하는 함수 및 말단 변수가 주어지면, 특징 공간 F에서 데이타를 구분하는 결정면들이 존재하게 된다. 이들 결정면들로부터 구분되는 공간은 버전공간(version space)이라고 부르며[31], 이 공간에는 "$f(x_i) \cdot y_i > 0$"을 만족하는 가설 f가 존재한다. 유전자 프로그래밍을 이용한 분류는 주어진 함수와 말단 변수를 이용하여 모든 샘플 x에 대해 "$f(x) \cdot y > 0$"을 만족하는 가설 f를 탐색하는 문제로 간주할 수 있다.

정의 1. 함수: $Fu = \{fu | fu \in \{+, -, \times, \div\}\}$

정의 2. 말단 변수: $TS = \{ts | ts \in Fe\}$, 단, Fe는 특징집합 $\{f_1, f_2, ..., f_n\}$

정의 3. 가설: $H = \{f | f(x_i) = t_{(fu,\ ts,\ depth)}(x_i)$, 단, $t \in T$ 이며 T는 F에 대응하는 트리 공간$\}$

정의 4. 버전공간: $V = \{f \in H \mid \forall i \in \{1, ..., m\}\ y_i \cdot f(x_i) > 0\}$, 단, m은 학습 샘플의 수

정의 5. 재정의된 버전공간: $V' = \{t \in T \mid y_i \cdot (t_{(fu,\ ts,\ depth)}(x_i)) > 0,\ i = 1, ..., m\}$

가설 집합 H가 존재하면 트리로 표현되는 규칙 t와 H의 가설 f는 일대일 대응이며, 버전공간은 특징 공간에서 학습 샘플을 선형으로 분리 가능할 때 존재한다. 특징 공간 F와 트리 공간 T 사이의 이중성 때문에, T에서의 점들은 F에서의 초평면에 대응한다. 즉, 특징 공간에서 학습 샘플 x_i가 관측되면, 초평면 집합은 x_i를 정확히 분류해야만 한다.

정의 6. 앙상블 가설: $EH = \{eh \mid eh(x_i) = Majority\_vote(t_1(x_i), ..., t_m(x_i))$, 단, $t_j \in T$, m은 개별 분류기 수$\}$

정의 7. 버전공간 V'의 부피: $Vol(t)$는 t를 만족하는 버전공간 V'의 부피

정의 8. t의 정확율

$$Acc(x_i) = \frac{Vol(y_i \cdot t_{j(fu,ts,depth)}(x_i) > 0)}{Vol(y_i \cdot t_{j(fu,ts,depth)}(x_i) > 0) + Vol(y_i \cdot t_{j(fu,ts,depth)}(x_i) < 0)} \quad (4)$$

만약 클래스 레이블이 y_i인 테스트 샘플 x_i가 주어지면 규칙 t_j는 $t_{j(fu,\ ts,\ depth)}(x_i)$를 출력한다. 결합 전략으로 다수결 투표를 고려할 때, eh는 $y_i \cdot t_{j(fu,\ ts,\ depth)}(x_i) > 0$인 t_j가 $m/2$개 이상 존재하면 정확히 분류를 수행하게 된다. 각 $y_i \cdot t_{j(fu,\ ts,\ depth)}(x_i) > 0$는 T에서 절반의 공간을 의미하고, $t_{j(fu,\ ts,\ depth)}(x_i) = 0$은 트리공간 T에서 버전공간 V'에 대해 결정면 역할의 초평면을 의미한다. 즉, $m/2$ t_j가 50%이상의 정분류율을 가진다면 앙상블 가설은 정확한 분류를 수행하게 된다.

정의 9. t_i와 t_j 사이의 교차:

$IntSec(t_i, t_j) = Vol(t_i) \cap Vol(t_j)$

정의 10. t_i와 t_j 사이의 합:

$Union(t_i, t_j) = Vol(t_i) \cup Vol(t_j)$

정의 11. t_i와 t_j 사이의 다양성:

$D(t_i, t_j) = Union(t_i, t_j) - IntSec(t_i, t_j)$

$\approx Distance(fu_i, fu_j) + Distance(ts_i, ts_j) + Distance(depth_i, depth_j)$

t_i가 이상적인 가설 t_{ideal}의 부분집합이라고 가정할 때, 무한한 수의 t_i를 결합한 앙상블 가설은 다음과 같이 이상적인 가설에 근접하게 된다.

$$\lim_{i \to \infty} Union(t_1, t_2, ..., t_i) \approx Vol(t_{ideal}) \quad (5)$$

그러나 실제로는 유한한 수의 t_i만이 결합가능하며, 이들 가설 사이의 교차영역이 최소화된다면 앙상블의 효과를 극대화시킬 수 있다. 결국, 트리 공간에서의 규칙의 다양성을 증가시키면 분류 성능을 향상시킬 수 있다. 다음의 수식과 같이 동일한 규칙을 결합한 경우보다 다른 규칙을 결합한 경우에 전체적으로 버전공간의 부피가 증가한다.

$$Union(t_i, t_i) = Vol(t_i) \leq Union(t_i, t_j), \text{ 단, } t_i \neq t_j \quad (6)$$

다시 말해서, 다양한 분류기를 결합한 경우가 버전공간에서 더 큰 부피를 가지며, 결과적으로 분류율의 향상을 가져온다.

$$D(eh_i) < D(eh_j),$$
$$\text{단, } eh_i \text{ 는 } m\text{개의 동일한 } t_i \text{ 의 앙상블, } eh_j \text{ 는 서로 다른 } t_i \text{ 의 앙상블 } (0 \leq i \leq m)$$
$$Vol(eh_i) = Vol(t_i) \leq Vol(eh_j) \approx Vol(t_{ideal}) \quad (7)$$

3. 실험 및 결과

3.1 실험 환경

제안하는 방법을 평가하기 위해 림포종 암 데이타[32], 폐 암 데이타[33], 난소 암 데이타[34] 등의 유전 발현 데이타 집합을 사용하였다. 모든 특징값은 0에서 1로 정규화하였다. 림프종 암(DLBCL)은 비홉킨스 림프종의 대표적인 질병이다. 림포마 암에는 다양한 종류가 있으며 각각 다른 치료방법이 필요하지만 임상적으로 이들을 구분하는 것은 쉽지 않다. 이 데이타 집합은 각 샘플이 4,026개의 유전 발현값으로 이루어진 47개의 샘플로 구성되어 있다. 24개 샘플은 germinal centre B-like group이고 23개 샘플은 activated B-like group이다.

폐 암 데이타 집합은 malignant pleural mesothelioma (MPM)와 adenocarcinoma (ADCA)를 분류하는 것이 목적이며, 각각 31개와 150개로 총 181개의 샘플로 이루어져 있다. 각 샘플은 12,533개의 유전 발현값으로 구성된다. 난소 암 데이타 집합은 혈청의 단백질 패턴을 분석하여 난소암을 구분하는 것이 목적이며, 91개의 정상인 샘플과 162개의 난소암 샘플로 구성되어 있다. 각 샘플은 15,154개의 유전 발현값으로 구성되어 있다.

각 데이타 집합은 많은 특징에 비해 샘플 수는 매우 적기 때문에 대표적인 5-fold cross-validation을 수행하여 평가하였다. 각 데이타 집합의 1/5은 테스트 데이타로, 나머지는 학습 데이타로 사용하며 모든 데이타가 한번 씩 테스트 데이타로 사용되도록 5회 실험한다. 이런 실험을 유전자 프로그래밍의 초기 집단을 바꿔가며 총 10회 반복하여 50(5×10)회의 실험 결과의 평균을 사용하였다. 유전자 프로그래밍에 사용되는 변수의 값은 표 1에서와 같이 설정하였다. 선택 방법으로는 엘리트 유지전략을 이용하는 룰렛휠 방법을 사용하였다.

표 1 실험 파라메터

파라메터	값
집단 크기	200
최대 세대수	3,000
선택율	0.6~0.8
교차율	0.6~0.8
돌연변이율	0.1~0.3
치환율	0.1
트리 최대 깊이	5
엘리트 유지전략	Yes

3.2 정분류율

앙상블 분류기의 성능을 평가하기 전에 유전자 프로그래밍의 분류기로서의 성능을 먼저 분석하였다. 기존에 많이 사용되는 기계 학습 분류기인 신경망과 결정 트리에 대해 림프종 암 분류에 대한 성능을 비교하였다. 성능분석은 LOOCV(Leave-one-out cross validation)를 따라 10회 실험을 반복하였다. 표 2는 림프종 암에 대한 대표적인 분류기의 분류 결과를 보여주는데, 유전자 프로그래밍의 성능이 매우 우수한 것을 확인하였다.

표 2 대표적 분류기를 이용한 림프종 암 분류결과

분류기	정분류율
신경망	94.6%(±3.5)
결정트리 (C4.5)	78.8%(±2.7)
결정트리 (See5)	82.3%(±2.5)

표 3은 각 암 데이타에 대한 분류 결과를 보여준다. 다중 분류기를 결합한 앙상블 분류기가 개별 분류기보다 약 1~10%의 성능 향상을 보이고 있다. 제안하는 방법은 다양성을 고려하지 않고 단순히 Bagging을 이용하여 10개의 분류 규칙을 결합한 경우와 동일 수인 5개의 분류 규칙을 결합한 경우보다 더 나은 성능을 보였다. 이는 다양성을 고려할 때 앙상블의 성능을 향상시킬 수 있음을 의미한다.

10개의 분류 규칙을 결합한 경우보다 단순히 5개의 분류 규칙을 결합한 경우가 더 나은 결과를 보였는데, 이는 개별 분류기의 수가 늘어남에 따라 오류도 함께 증가하기 때문이다. 결과적으로 제안하는 방법은 10개의 분류 규칙을 사용할 때와 유사한 정보량을 유지하면서 오류는 최소화하는 효과를 보였다. 결합 전략의 측면에서는 MAJ가 DT보다 약간 나은 성능을 보였으며, 두 방법 모두 개별 분류기의 성능보다 많이 향상된 결과를 얻을 수 있었다. 림프종 암 분류의 경우 분류 규칙 5개를 결합한 경우와 제안한 방법 사이의 유의성 평가를

표 3 분류 결과

데이타	결합방법	제안하는 방법	5개 단순결합	10개 단순결합	개별분류기
림프종 암	MAJ	98%(±4.6)	93.7%(±9.7)	92.2%(±11.7)	91.3%(±11.8)
	DT	97.4%(±4.6)	94.6%(±9.3)	95.6%(±7.8)	
폐 암	MAJ	99.5%(±1.1)	99.1%(±1.3)	99.2%(±1.3)	97.8%(±2.1)
	DT	99.2%(±1.9)	98.7%(±2.2)	99.2%(±1.9)	
난소 암	MAJ	91.2%(±2.9)	89.2%(±3.7)	89.3%(±3.2)	87.7%(±5.2)
	DT	87.9%(±5.1)	87.6%(±4.8)	88.8%(±4.8)	

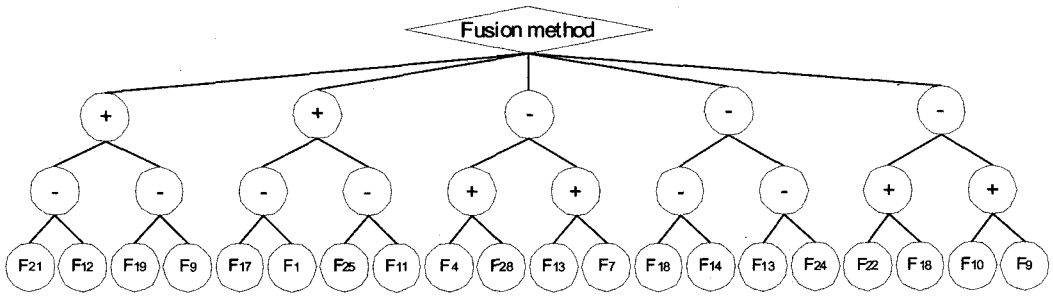

그림 4 획득된 앙상블 분류기의 모습

수행한 결과, 유의수준 5%에서 -43.1인 z-통계치를 얻어 제안한 방법의 유용성을 확인하였다. 그림 4는 획득된 분류규칙의 모습을 보여준다.

3.3 다양성에 따른 성능 분석

본 절에서는 먼저 그림 5에서와 같이 다양성과 분류 성능의 관계를 분석하였다. 모든 결합 전략에서 분류 규칙 사이의 다양성이 증가함에 따라 분류 성능이 향상되는 것을 확인하였다. 그림 5는 림프종 암 데이타에 대한 다양성과 분류 성능의 관계를 보여주며, 다른 데이타에서도 비슷한 양상을 확인하였다.

기존의 다양성 측정 방법인 Q-통계, 상관계수, 불일치 척도(disagreement measure), 공동 오류 척도(double-fault measure), Kohavi-Wolpert 차이, 사정동의(measurement of interrater agreement), 엔트로피, 난이성, 범용 다양성(generalized diversity)과 동시오류 다양성(coincident failure diversity)을 대상으로 비교 실험을 수행하였다. 이들 방법에 대한 자세한 설명은 [19,27]에 있다.

기존의 다양성 측정 방법은 정확한 다양성 측정을 위해서는 보다 많은 수의 샘플이 필요하며, 분류 성능이 높을 경우에는 성능향상이 저조하다. 이는 앞서 설명한 정확성-다양성 딜레마에 의한 것이다. 학습 데이타가 적은 유전자 발현 데이타의 경우, 모든 학습 데이타를 정확히 분류하는 규칙이 획득되어 그들 사이의 다양성이 매우 적거나 0이 되었다. 이와는 달리, 제안하는 방법은 분류 규칙의 구조 자체를 분석하였기 때문에 높은 성능을 갖는 분류 규칙들에 대해서도 다양성 측정이 가능하였고 분류 성능도 향상시킬 수 있었다. 그림 6은 림프종 암 데이타에 대한 다양성 측정 방법의 성능 비교를 보여준다. 기존의 방법으로는 다양성에 따른 성능향상이 없었지만 제안하는 방법은 분류 규칙의 구조적 다양성을 측정하여 학습 데이타에 대한 분류 규칙의 성능과는 상관없이 성능향상을 획득하였다. 폐 암과 난소 암 데이타에 대해서도 비슷한 양상을 확인하였다.

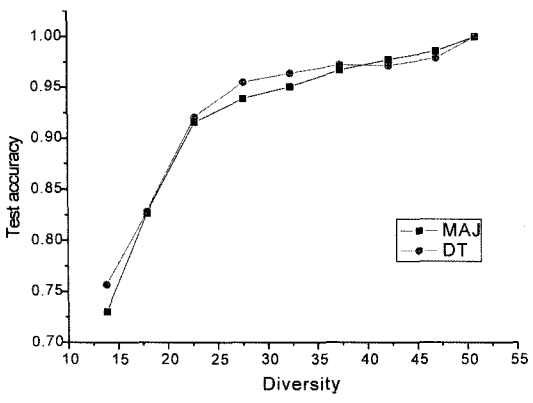

그림 5 다양성에 따른 분류 성능 향상 (림프종 암 데이타)

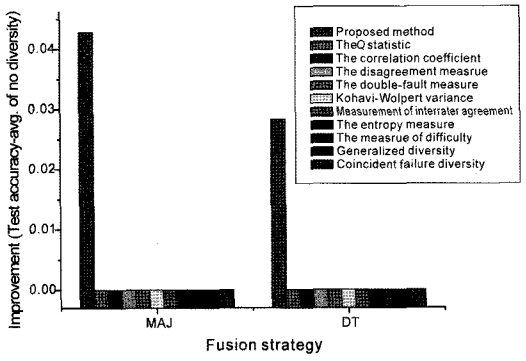

그림 6 기존 다양성 측정 방법과의 성능 비교(림프종 암 데이타)

4. 결 론

본 논문에서는 유전자 프로그래밍을 위한 효과적인 앙상블 방법을 제안하였다. 유전 발현 데이타는 많은 수의 특징으로 구성되어 있기 때문에 특징 선택 과정을 통해 특징의 차원을 축소하였고, 유전자 프로그래밍을 이용하여 다수의 암 분류 규칙을 생성하였다. 기존의 앙상블 방법과는 달리 분류 규칙 사이의 다양성을 직접

측정하여 일부 분류 규칙을 선택하였고 이들을 결합기법을 통해 결합하였다. 유전자 프로그래밍은 해석에 용이한 분류 규칙의 생성에 효과적이며, 직접적인 규칙 비교를 통한 다양성의 측정을 가능하게 한다.

유전 발현 데이타를 이용한 암 분류 문제에 제안하는 방법을 적용하였고, 그 유용성을 확인하였다. 단일 분류 규칙이나 10개 또는 5개의 분류 규칙을 단순히 결합한 경우보다 다양성을 고려하여 분류 규칙을 선택한 후 결합한 경우가 보다 높은 성능을 보여주었다. 특히, 다양성이 증가할 때 분류 성능도 함께 증가하는 것을 확인할 수 있었을 뿐만 아니라 제안하는 방법의 우위성도 확인하였다. 향후에는 생물 정보학 분야의 다양한 데이타에 제안하는 방법을 적용하여 성능을 보다 폭넓게 평가할 것이다.

참 고 문 헌

[1] U. Schmidt and C. Begley, "Cancer diagnosis and microarrays," *The Int. J. of Biochemistry & Cell Biology*, vol. 35, no. 2, pp. 119-124, 2003.

[2] I. Sarkar, et al., "Characteristic attributes in cancer microarrays," *J. of Biomedical Informatics*, vol. 35, no. 2, pp. 111-122, 2002.

[3] J. Khan, et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[4] V. Roth and T. Lange, "Bayesian class discovery in microarray datasets," *IEEE Trans. Biomedical Engineering*, vol. 51, no. 5, pp. 707-718, 2004.

[5] C. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349-358, 2001.

[6] N. Camp and M. Slattery, "Classification tree analysis: A statistical tool to investigate risk factor interactions with an example for colon cancer," *Cancer Causes and Control*, vol. 13, no. 9, pp. 813-823, 2002.

[7] L. Li, et al., "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.

[8] J. Deutsch, "Evolutionary algorithms for finding optimal gene sets in microarray prediction," *Bioinformatics*, vol. 19, no. 1, pp. 45-52, 2003.

[9] M. Karzynski, et al., "Using a genetic algorithm and a perceptron for feature selection and supervised class learning in DNA microarray data," *Artificial Intelligence Review*, vol. 20, no. 1-2, pp. 39-51, 2003.

[10] W. Langdon and B. Buxton, "Genetic programming for mining DNA chip data for cancer patients," *Genetic Programming and Evolvable Machines*, vol. 5, no. 3, pp. 251-257, 2004.

[11] G. Valentini, "Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles," *Artificial Intelligence in Medicine*, vol. 26, no. 3, pp. 281-304, 2002.

[12] C. Park and S.-B. Cho, "Evolutionary computation for optimal ensemble classifier in lymphoma cancer classification," *Lecture Notes in Artificial Intelligence*, vol. 2871, pp. 521-530, 2003.

[13] A. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," *Applied Bioinformatics*, vol. 2, no. 3 Suppl., pp. S75-S83, 2003.

[14] L. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281-286, 2002.

[15] R. Bryll, et al., "Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets," *Pattern Recognition*, vol. 36, no. 6, pp. 1291-1302, 2003.

[16] G. Webb and Z. Zheng, "Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 8, pp. 980-991, 2004.

[17] D. Optiz and R. Maclin, "Popular ensemble methods: An empirical study," *J. of Artificial Intelligence Research*, vol. 11, pp. 169-198, 1999.

[18] M. Islam, et al., "A constructive algorithm for training cooperative neural network ensembles," *IEEE Trans. Neural Network*, vol. 14, no. 4, pp. 820-834, 2003.

[19] C. Shipp and L. Kuncheva, "Relationships between combination methods and measures of diversity in combining classifiers," *Information Fusion*, vol. 3, no. 2, pp. 135-148, 2002.

[20] J.-H. Hong and S.-B. Cho, "Rule discovery for cancer classification using genetic programming based on arithmetic operators," *J. of Korea Information Science Society: Software and Applications*, vol. 31, no. 8, pp. 999-1009, 2004.

[21] J. Koza, "Genetic programming," *Encyclopedia of Computer Science and Technology*, vol. 39, pp. 29-43, 1999.

[22] Y. Zhang and S. Bhattacharyya, "Genetic programming in classifying large-scale data: An ensemble method," *Information Sciences*, vol. 163, no. 1-3, pp. 85-101, 2004.

[23] M. Brameier and W. Banzhaf, "Evolving teams of predictors with linear genetic programming," *Genetic Programming and Evolvable Machines*, vol. 2, no. 4, pp. 381-407, 2001.

[24] F. Fernaandez, et al., "An empirical study of multipopulation genetic programming," *Genetic Programming and Evolvable Machines*, vol. 4, no. 1, pp. 21-51, 2003.
[25] K. Imamura, et al., "Behavioral diversity and a probabilistically optimal GP ensemble," *Genetic Programming and Evolvable Machines*, vol. 4, no. 3, pp. 235-253, 2003.
[26] G. Zenobi and P. Cunningham, "Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error," *Lecture Notes in Computer Science*, vol. 2167, pp. 576-587, 2001.
[27] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181-207, 2003.
[28] T. Windeatt, "Diversity measures for multiple classifier system analysis and design," *Information Fusion*, 2004.
[29] E. Bruke, et al., "Diversity in genetic programming: An analysis of measures and correlation with fitness," *IEEE Trans. Evolutionary Computation*, vol. 8, no. 1, pp. 47-62, 2004.
[30] L. Kuncheva, et al., "Decision templates for multiple classifier fusion: An experimental comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299-314, 2001.
[31] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. of Machine Learning Research*, vol. 2, pp. 45-66, 2001.
[32] A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503-511, 2000.
[33] G. Gordon, et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, no. 17, pp. 4963-4967, 2002.
[34] E. Petricoin III, et al., "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, no. 9306, pp. 572-577, 2002.

홍 진 혁
정보과학회논문지 : 소프트웨어 및 응용
제 32 권 제 11 호 참조

조 성 배
정보과학회논문지 : 소프트웨어 및 응용
제 32 권 제 11 호 참조