유전알고리즘을 이용한 유전자발현 데이타상의 특징-분류기쌍 최적 앙상블 탐색

(Searching for Optimal Ensemble of Feature-classifier Pairs in Gene Expression Profile using Genetic Algorithm)

박 찬 호 † 조 성 배 ††
(Chanho Park) (Sung-Bae Cho)

요 약 유전발현 데이타는 생명체의 특정 조직에서 채취한 샘플을 microarray상에서 측정한 것으로, 유전자들의 발현 정도가 수치로 나타난 데이타이다. 일반적으로 정상조직과 이상조직에서 관련 유전자들의 발현정도는 차이를 보이기 때문에, 유전발현 데이타를 통하여 질병을 분류할 수 있다. 이러한 분류에 모든 유전자들이 관여하지는 않으므로 관련 유전자를 선별하는 작업인 특징선택이 필요하며, 선택된 유전자들을 적절히 분류하는 방법이 필요하다. 본 논문에서는 상관계수, 유사도, 정보이론 등에 기반을 둔 7가지 특징 선택 방법과 대표적인 6가지 분류기에 대하여 특징-분류기 쌍의 최적 앙상블을 탐색하기 위한 유전자 알고리즘 기반 방법을 제안한다. 두 가지 암 관련 유전자 발현 데이타에 대하여 leave-one-out cross validation을 포함한 실험을 해본 결과, 림프종 데이타와 대장암 데이타 모두 단일 특징-분류기 쌍보다 훨씬 우수한 성능을 보이는 앙상블들을 발견할 수 있었다.

키워드 : 유전발현 데이타, 특징선택, 분류기, 앙상블, GA

Abstract Gene expression profile is numerical data of gene expression level from organism, measured on the microarray. Generally, each specific tissue indicates different expression levels in related genes, so that we can classify disease with gene expression profile. Because all genes are not related to disease, it is needed to select related genes that is called feature selection, and it is needed to classify selected genes properly. This paper proposes GA based method for searching optimal ensemble of feature-classifier pairs that are composed with seven feature selection methods based on correlation, similarity, and information theory, and six representative classifiers. In experimental results with leave-one-out cross validation on two gene expression profiles related to cancers, we can find ensembles that produce much superior to all individual feature-classifier fairs for Lymphoma dataset and Colon dataset.

Key words : gene expression profile, feature selection, classifier, ensemble, GA

1. 서 론

지난 몇 년간 암의 조기 발견과 정확한 분류를 위한 연구가 활발하게 진행되어 왔지만, 아직 완벽한 방법을 제시한 연구는 없었다. 이는 암의 원인이 되는 경로가 다양할 뿐 아니라, 매우 많은 변이가 존재하며, 실험적으로 대량의 데이타를 얻기가 힘들었기 때문이다. 그러나 최근의 생명공학 및 분석화학과 관련된 DNA microarray기술의 급격한 발달은 생명체에 관한 대량의 유전정보를 얻는 것을 가능하게 해주었다. 이렇게 얻어진 유전정보의 원시형태는 단순한 숫자들의 나열이기 때문에, 직접 그 의미를 발견하기는 힘들다. 따라서 이것을 분석하기 위하여 수 년 전부터 많은 분석 방법이 연구되어 왔으며, 현재도 많은 그룹에서 연구가 진행 중이다[1,2].

유전발현 데이타의 분류를 위해서 먼저 할 일은 실험을 통하여 얻은 데이타를 정규화 시키는 것이다[3]. 정규화를 통하여 얻어진 데이타로부터 의미 있는 유전자들을 선택하는 작업인 특징 선택 작업을 하게 되는데, 이는 일반적으로 microarray를 통해서 나오는 데이타는

· 본 연구는 보건복지부 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임
† 학생회원 : 연세대학교 컴퓨터과학과
cpark@sclab.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@cs.yonsei.ac.kr
논문접수 : 2003년 4월 3일
심사완료 : 2003년 12월 5일

샘플의 수에 비하여 유전자의 수가 매우 많고, 그 중 상당수는 데이타의 분류에 있어 도움을 주지 못하기 때문이다[4]. 특징선택 단계를 통하여 선택된 유전자들은 분류기의 입력으로 들어가는데, 분류기는 학습 집단의 유전자 발현 패턴을 입력으로 받아 입력패턴이 최대한 바른 출력을 내도록 학습이 이루어진다. 이렇게 학습된 분류기는 평가집단이나 테스트 집단에 대하여 실제로 얼마나 정확한지 평가받는다[5,6].

한편 완벽한 특징-분류기를 찾기는 매우 어렵기 때문에 분류에 앙상블 분류기를 이용한 방법이 시도되어 왔다[7]. 앙상블이란 특징-분류기 쌍을 결합하는 것을 말하는데, 앙상블을 통하여 안정적이고 좋은 결과를 얻을 수 있다는 것이 알려져 있으며, 넓은 해 공간을 얻을 수 있다. 앙상블을 하는 방법은 여러 가지가 있는데, 대표적으로 투표, 가중투표, 가중평균, 베이지안 결합 등이 있다. 유전발현 데이타의 분류문제도 마찬가지로 앙상블을 이용하여 좋은 성능을 얻을 수 있다. 반면 앙상블은 그 시간이 오래 걸린다는 단점이 있다. 특징-분류기를 하나 사용하는 것부터 모두 이용하여 결합하는 것까지 엄청난 수의 앙상블이 존재하기 때문이다. 만약 m개의 특징선택 방법과 n개의 분류기가 있다면 mn개의 특징-분류기 쌍이 나오게 된다. 본 논문에서는 7개의 특징선택 방법과 6개의 분류기를 이용하였으므로 42개의 특징-분류기 쌍이 생성되었으며, 이들을 대상으로 한 가능한 앙상블의 수는 다음 식과 같이 표현된다(k는 앙상블에 사용되는 특징-분류기의 수).

$$\sum_{k=1}^{42} {}_{42}C_k \cong 4\times 10^{12} \quad (1)$$

이 중에 어떤 것이 최적인지 알아보기 위하여 모든 앙상블을 비교할 수 하지만, 이는 거의 불가능에 가까울 정도로 매우 많은 시간이 소요된다. 그리고 만약 이 상태에서 단 하나의 특징선택 방법이나 분류기가 추가된다고 하여도 연산시간은 지수적으로 증가기 때문에 모든 결과를 구한다는 것은 더욱 불가능해진다. 따라서 효과적으로 최적의 앙상블을 탐색하는 방법이 필요하며, 본 논문에서는 유전자 알고리즘(genetic algorithm, GA)에 기반을 둔 최적의 특징-분류기 쌍 앙상블을 탐색 방법을 제안한다. GA의 초기 염색체 집단을 임의로 생성하여, 그들이 유전연산을 통하여 최적 앙상블에 접근하는 양상을 보이는지 알아본다. 아울러 두 가지 암 관련 데이타인 림프종 데이타와 대장암 데이타에 대하여 적용시켜 보고, 제안한 방법의 타당성을 평가한다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 2장은 연구의 배경이 되는 DNA microarray기술과 GA에 대하여 소개하고, 관련연구에 대하여 기술한다. 3장은 제안하는 시스템의 구성과 각 세부사항에 대하여 설명한다. 4장은 실험 과정과 결과를 설명하고, 5장은 논문의 결론과 향후연구에 대해 언급한다.

2. 배경

이 장에서는 분석의 대상이 되는 유전발현 데이타를 얻기 위한 DNA microarray기술과 연구의 배경이 되는 GA에 대하여 설명하고, 관련연구에 대하여 소개한다.

2.1 DNA microarray

DNA microarray는 용액이 투과되지 않는 딱딱한 지지체 위에 고밀도로 cDNA를 고정시켜 놓은 것으로 DNA chip이라고도 한다. Array상의 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 각각 다른 형광물질을 합성한 것을 동일한 양으로 보합한 것이다. 이것을 레이저 형광 스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현정도를 얻을 수 있는데, Cy5/Cy3의 비율에 밑이 2인 로그를 취한 값을 그 셀의 발현정보 값으로 얻게 된다[8].

$$gene\_\exp ression = \log_2 \frac{Int(Cy5)}{Int(Cy3)} \quad (2)$$

데이타를 얻는 전체적인 과정은 그림 1과 같다. 그림에서 각 직사각형의 판이 하나의 array에 해당하고, array 안에서 격자 모양으로 박혀 있는 것이 array의 각 셀이 되며, 각 셀에는 대응되는 유전자들이 심어져있다. 이 array위로 형광물질이 처리된 test와 reference에 해당하는 유전물질들이 해당 셀에 달라붙게 되고, 그것들이 셀에 달라붙은 정도에 따라서 다른 형광 정도를 보여준다.

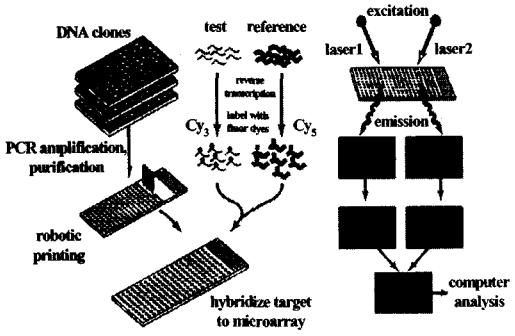

그림 1 DNA microarray 데이타 취득 과정

하나의 microarray를 통하여 전체 지놈(genome)을 탐색할 수 있고, 동시에 수천 개 유전자간의 상호관계를 분석할 있는 가능성도 제공해주기 때문에, 다량의 유전정보를 얻기 위해서는 microarray의 이용이 필수적이다.

2.2 유전자 알고리즘

유전자 알고리즘(genetic algorithm, GA)은 생명체의

유전 및 진화과정을 전산학적으로 모델링한 기계학습 방법으로, 탐색해야 할 공간이 매우 넓은 경우 유용하게 사용되는 탐색 및 최적화 기법이다[9]. GA의 기본 요소는 염색체(chromosome)라 불리는 것으로서 이들이 집단을 이루어 해를 탐색한다. 각 염색체는 적자생존의 법칙에 의하여 상대적으로 우수한 것이 살아남을 확률이 크며, 또한 유전연산자에 의하여 진화과정을 거치게 된다. 한 세대를 이루는 염색체들이 진화하여 다름 세대로 되는 과정은 그림 2와 같다.

먼저 초기 집단을 이루는 염색체들이 임의로 선택된다. 이들은 첫 세대의 부모집단이 되어 적합도를 평가받은 후, 그 값을 기반으로 하여 확률적으로 선택된다. 상대적으로 높은 적합도를 받은 염색체는 여러 번 선택이 되며, 낮은 것들은 선택되지 않기도 한다. 선택된 염색체들은 두개씩 짝을 지어 교차연산을 거치게 되고, 마지막으로 돌연변이연산을 통해 새로운 세대의 집단이 생성된다. 교차연산에서는 짝지어진 염색체끼리 유전정보를 교환하는 작업이 일어나며, 돌연변이연산에서는 유전정보 중 일부가 사라지거나 새로 나타난다. 새롭게 생성된 집단은 그들의 부모세대가 했던 과정을 반복하게 되며 이는 미리 정의한 수준에 이를 때까지 반복된다.

만약 교차나 돌연변이로 인하여 새롭게 생긴 염색체가 적합도를 높여주는 역할을 한다면 그 염색체는 살아남을 확률이 커서 자손을 널리 퍼뜨릴 가능성이 높아진다. 그렇지 않다면 그 염색체는 다음번 선택과정에서 곧 도태된다. 이 사실로부터 GA 진화의 방향이 염색체들의 평균 적합도가 증가하는 쪽이라는 것을 알 수 있다.

2.3 관련연구

DNA microarray데이타를 분류하기 위한 연구가 많은 그룹에서 진행 중이다. 그들이 의미 있는 유전자를 뽑기 위하여 사용한 특징선택 방법으로는 정보이득(information gain), 신호 대 잡음비(signal to noise ratio), t-통계량, 피어슨 상관계수, 주성분 분석(principal component analysis)등이 있다[3,4,10-12]. 또한 분류기로는 다층신경망, k-최근접 이웃, 결정 나무(decision tree), SVM, 피셔의 선형판별식(Fisher's linear discriminant analysis)등을 사용하였다[4-6,13,14]. 표 1에 관련 연구가 요약되어 있다.

한편 분류의 성능을 높이기 위하여 앙상블에 관한 많은 연구가 진행 중이다. 앙상블은 분류의 성능을 높여주기도 하지만 개별 분류기의 결과가 편차가 심한데 비해서 안정된 결과를 내놓는다[7,15]. 앙상블에 이용되는 분류기들이 서로 독립적이고, 각 분류기의 인식률이 50%를 넘길 경우 앙상블에 참여하는 분류기가 늘어날수록 앙상블이 좋은 성능을 내는 것으로 알려져 있다.

3. 최적의 특징-분류기 쌍 앙상블 분류기

본 논문에서 제안한 유전발현 데이타 분류구조의 전체적인 흐름은 그림 3과 같다. 먼저 microarray를 통해서 얻은 데이타를 정규화 시킨 후 특징선택기의 입력으

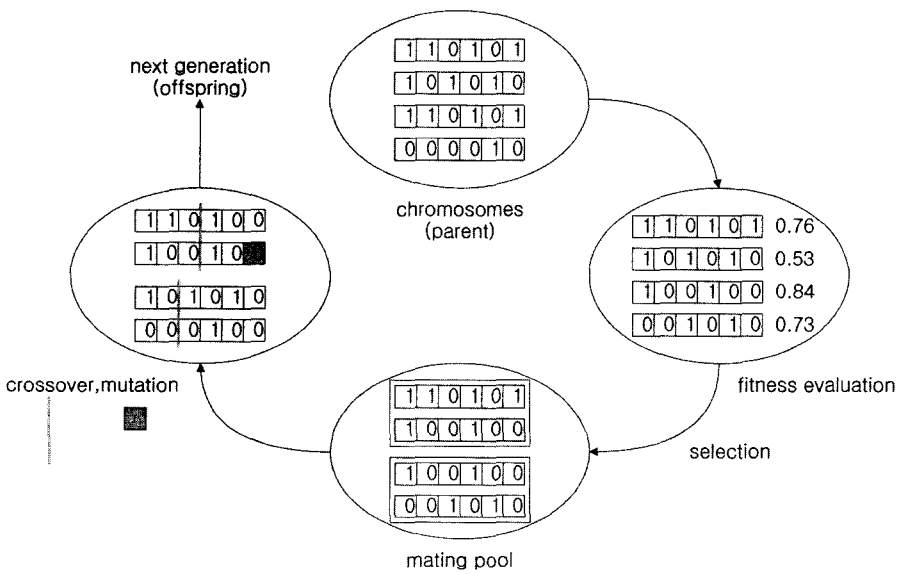

그림 2 GA의 한 세대 진화

표 1 DNA microarray 분류 관련 연구들

저자	데이타	사용한 방법		인식률(%)
		특징선택방법	분류기	
Furey 등	백혈병	Signal to noise ratio	SVM	94.1
	대장암			90.3
Li 등	림프종	Genetic algorithm	KNN(training set/test set 동일)	84.6~
	대장암			88.2~
Dudoit 등	백혈병	The ratio of between-groups to within-groups sum of squares	Nearest neighbor	95.0~
	림프종			95.0~
	백혈병		Diagonal linear discriminant analysis	95.0~
	림프종			95.0~
Nguyen 등	백혈병	Principal component analysis	Logistic discriminant	94.2
	림프종			98.1
	대장암		Diagonal linear discriminant analysis	87.1
	백혈병		Quadratic discriminant analysis	95.4
	림프종		Boost CART	97.6
	대장암			87.1

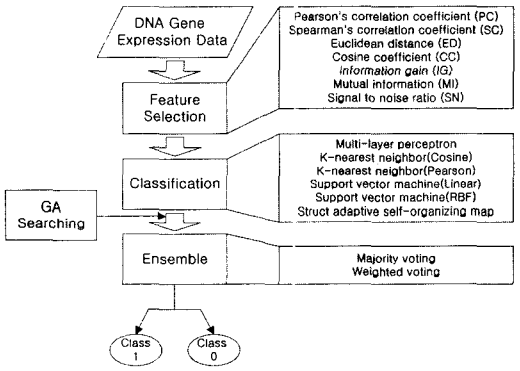

그림 3 앙상블 탐색 시스템

로 넣고, 각 특징선택 방법에 의하여 의미있는 유전자들을 선택한다. 선택된 유전자들은 분류기의 입력으로 들어가 결과를 내어 놓는다. 특징선택 방법 및 분류기에 따라 여러 개의 특징-분류기가 만들어지며 이들 각각은 앙상블의 대상이 된다. GA는 수많은 앙상블 중에서 최적의 것을 탐색해 나간다.

3.1 유전자 선택

유전자 선택과정에서는 분류에 도움을 줄 것이라 예상되는 유전자들을 선택하는 작업을 한다. 보통 유전발현 데이타는 샘플 수가 수십 개 정도밖에 되지 않는데 그것을 설명하기 위한 유전자들은 보통 수천 개 이상이기 때문에 그들을 모두 이용하는 것이 계산상 힘들 뿐 아니라, 그 중 상당수는 분류에 도움을 주지 못하기 때문에 유전자를 선택하는 작업을 먼저 한다. 본 논문에서는 유전자 선택의 기준으로 상관계수, 유사도, 정보이론에 기반을 둔 일곱 가지 특징선택 방법을 사용하였다. 먼저 각 선택 방법 별로 점수를 계산하여, 그 중 상위에 랭크된 유전자들을 선택하여 분류기의 입력으로 사용하였다[7].

(1) 상관계수 기반 방법

상관계수분석이란 변수간의 관련성을 분석하기 위해 사용하는 방법으로서, 하나의 변수가 다른 변수와 관련성이 있는지, 또 관련 정도가 어느 정도인지 알아보기 위한 방법이다. 피어슨 상관계수는 상관계수분석에서 자주 이용되는 계수이며, 상관계수 r은 [-1, 1]의 값을 갖는다. r의 값이 1에 가까울수록 두 변수는 양의 상관관계를 나타내고, 서로 유사하다는 것을 의미한다. 반면 r이 -1에 가깝다면 두 변수의 관계는 음의 상관관계가 되며 서로 반대의 작용을 하는 관계가 있다는 것을 의미한다. r이 0에 가깝다면 두 변수 사이에 별로 관계가 없음을 의미한다. N개의 원소를 갖는 두 벡터 X와 Y사이의 피어슨 상관계수(PC)는 식 (3)과 같이 정의된다.

$$r_{pearson} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (3)$$

한편 변수들의 값을 직접 이용하는 모수 분석과는 달리 양적 변수가 아니어도 될 때 이용할 수 있는 상관계수분석 방법이 존재하는데 이를 비모수 분석이라 하고, 스피어맨 상관계수 같은 것이 있다. 스피어맨 상관계수는 변수의 순위배열을 사용하여 변수간의 상관관계를 분석하는 방법으로 피어슨 상관계수와 마찬가지로 상관계수 r은 [-1, 1]의 값을 갖는다. 한편 스피어맨 상관계수(SC)는 X와 Y의 순위배열 Dx와 Dy를 사용하여 식 (4)와 같이 나타낼 수 있다.

$$r_{spearman} = 1 - \frac{6\sum(Dx - Dy)^2}{N(N^2 - 1)} \quad (4)$$

(2) 유사도 기반 방법

상관계수분석이 두 변수의 상관정도를 분석하는 방법이라면 유사도 측정법은 두 변수의 유사성을 측정하는 방법이다. 두 변수간의 유사성은 거리로 나타낼 수 있는데, 거리가 가까울수록 유사성이 높음을 뜻한다. 유클리드 거리는 두 변수간의 기하학적 공간에서의 거리를 나타내며, 거리 값이 크게 나올수록 유사한 정도가 낮은 것이기 때문에 그 값 자체로는 사실 비유사성 정도를 나타낸다고 볼 수 있다. 두 벡터 X와 Y의 유클리드 거리(ED)는 식 (5)와 같이 간단한 식으로 표현할 수 있다.

$$r_{euclidean} = \sqrt{\sum(X - Y)^2} \quad (5)$$

두 변수간의 유사성은 거리뿐만 아니라 두 변수사이의 각으로도 나타낼 수 있는데 각이 작을수록 같은 방향을 가리키며 서로 유사하다. 각을 직접 구하기보다는 그 각에 대한 코사인 값을 구하는 것이 쉬우므로 유사성은 코사인 값으로 나타낼 수 있다. 두 변수사이의 각이 작을수록 코사인 값은 1에 가깝고, 변수가 완전히 반대의 방향을 가리킬 때 코사인 값은 -1로 나타나므로 코사인 계수는 [-1, 1]의 값을 가지게 되고, 1에 가까울수록 유사성이 높다는 것을 의미한다. 두 변수 사이의 코사인 계수(CC)는 식 (6)과 같이 나타낼 수 있다.

$$r_{cosine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad (6)$$

(3) 정보이론 기반 방법

한편 전체 데이타로부터 의미 있는 정보를 뽑아내는 척도로 정보이론에서 사용하는 정보이득(Information Gain), 상보정보(Mutual Information), 신호 대 잡음 비(Signal to Noise Ratio) 등의 방법을 이용할 수 있다. 정보이득과 상호정보의 경우는 특정 유전자의 i번째 샘플이 특정 클래스 c에 속하는 가의 여부와 그 유전자가 발현했는가 여부의 두 가지 기준에 의하여 네 가지로 구분 짓고, 각 종류에 속하는 샘플의 수를 각각 A, B, C, D라 했을 때, 주어진 유전자 g의 정보 이득(IG)과 상호 정보 계수(MI)는 각각 식 (7), (8)과 같다.

$$IG = A \cdot \log\frac{A}{(A+B)\cdot(A+C)} + B \cdot \log\frac{B}{(A+B)\cdot(B+D)} \quad (7)$$

$$MI = \log\frac{A}{(A+B)\cdot(A+C)} \quad (8)$$

한편, 학습 샘플에 대해 주어진 유전자 g를 클래스 c에 속하는 것들과 그렇지 않은 것들로 분류한 후, 각각에 대하여 정규분포를 계산하였을 때, 클래스 c에 의하여 분류되는 유전자 g의 신호 대 잡음 비(SN)는 식 (9)와 같이 계산된다.

$$P(g,c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \quad (9)$$

일반적으로 선택되는 특징의 개수가 너무 적으면 정보를 제대로 포함하지 못하는 문제가 있고, 너무 많으면 노이즈가 많이 포함된다는 문제가 있다. 따라서 적절한 특징 개수의 선택이 중요한 문제가 되는데, 유전발현 데이타의 경우 20~70개 정도를 사용하였을 경우 큰 차이 없이 비슷한 성능이 나왔다. 따라서 통일성을 위하여 본 논문에서는 모두 25개씩의 유전자들을 사용하여 분류하였다.

3.2 분류기

분류기로는 패턴인식 분야에서 널리 사용 중인 다층 신경망, k-최근접 이웃, SVM 및 SOM을 수정한 SASOM을 사용하였다.

(1) 다층신경망

신경망은 인간의 신경회로를 모방하여 만든 예측모델이다. 신경망에도 은닉층의 존재여부, 학습 방식 등에 따라 여러 가지 종류가 존재하는데, 일반적으로 오류 역전파 알고리즘을 이용한 다층신경망(multi-layer perceptron, MLP)을 널리 사용한다. MLP의 구조는 입력층, 은닉층, 출력층으로 이루어진다[5].

신경망의 학습과정은 다음과 같다. 먼저 각 노드간의 에지(edge)는 임의의 가중치를 부여받는다. 그 후 입력이 들어오면 가중합을 계산하여 은닉노드로 전달시키고, 은닉노드에서는 그 값을 활성함수의 입력으로 하여 출력을 낸다. 은닉층의 모든 노드들이 출력을 내면 이는 다시 가중합이 계산되어져 출력층 활성함수의 입력으로 들어간다. 출력층은 활성함수의 출력을 구하고, 이를 목표값과 비교하여 오류를 계산한다. 그 후 오류를 줄이는 방향으로 가중치가 조절되는데 이때 오류 역전파(back propagation) 알고리즘이 사용된다[5,16].

(2) K-최근접 이웃

인공신경망이 학습 데이타들을 이용하여 분류기를 학습시키는데 시간이 많이 걸리는 단점이 있는 것에 반하여, K-최근접 이웃(k-Nearest Neighbor, kNN)은 분류기의 학습에 걸리는 시간이 거의 들지 않는다는 장점이 있다[9]. 대신 인공신경망이 분류기의 구조가 학습된 후에는 매우 빠른 속도로 테스트 샘플의 결과를 내주는데 반하여, kNN의 경우는 각 테스트 샘플의 결과를 내는 것에 학습 샘플의 수에 비례하는 만큼의 많은 시간이 걸리는 단점이 있다.

KNN의 동작원리는 간단하다. 테스트 샘플이 입력되면 이것과 각 학습 샘플과의 유사도를 계산하고 그 중 k개의 가장 가까운 학습 샘플을 선택한다. 들어온 테스트 샘플은 k개중 많은 수를 차지하는 집단에 속하는 것으로 판단되어 진다[17]. 일반적으로 이진 분류에 있어

서 애매함을 방지하기 위하여 k값은 홀수를 사용한다. KNN은 유사도를 측정하는 방법에 따라서 결과가 달라질 수 있으며, 특징선택방법에서 사용한 여러 가지 방법들이 유사도 측정도구로 사용될 수 있다. 본 논문에서는 피어슨 계수와 코사인 계수의 두 가지 유사도 측정도구를 사용하였다.

(3) SVM

SVM(Support Vector Machine)은 근래에 기계학습 분야에서 매우 각광받는 방법으로, 수학적 배경이 탄탄한 이진분류기이다. SVM은 샘플들이 선형 분리 경계를 가지고 분포해 있는 경우, 샘플들을 두 개의 클래스로 분리할 수 있는 최적의 초평면(hyperplane)을 찾아준다[6,18]. 여기서 최적의 초평면이란 두 집단에 속한 가장 가까운 샘플과의 거리가 최대가 되도록 하는 초평면을 뜻한다. 샘플들이 비선형 공간에 분포하고 있는 경우 커널함수를 사용하여 샘플 공간을 고차원의 선형 특징공간으로 사상시킨 후 최적의 초평면을 찾아주게 된다. 커널함수로는 선형 커널함수, 다항 커널함수, RBF 커널함수 등 여러 가지가 쓰일 수 있다. 본 논문에서는 웹상에 공개되어 있는 Joachims의 SVM light(http://svmlight.joachims.org/)를 이용하여, 선형과 RBF커널의 두 가지를 사용하였다.

(4) SASOM

Kohonen에 의하여 개발된 SOM(Self-Organizing Map)은 클러스터링(clustering) 방법 중 하나로, 입출력 공간의 위상을 보존하는 좋은 특성이 있다. 하지만 SOM은 학습이 시작되기 전에 지도의 구조를 결정해야 하는 단점이 있기 때문에 실제 분류문제에 직접 적용하기는 어렵다. SASOM(Structure Adaptive Self-Organizing Map)은 이러한 SOM의 단점을 보완하기 위해 제안된 방법이다. 이 방법은 기존의 SOM알고리즘을 이용하여 지도를 학습시킨 후, 학습된 지도의 노드들 중 서로 다른 집단의 데이타가 섞여있는 노드를 반복적으로 분화하여 주어진 데이타에 대하여 최적의 위상을 갖는 지도를 생성한다[19]. 본 논문에서 사용한 SASOM의 동작과정은 표 2와 같다.

표 2 SASOM의 동작과정

1단계. 지도를 4X4 크기로 초기화시킨 후 입력을 받아들인다.
2단계. SOM알고리즘을 사용하여 학습시킨다.
3단계. 적중률 95%미만의 노드를 찾아내어 그 노드를 2X2로 분화시킨다.
4단계. 분화된 노드를 LVQ알고리즘을 사용하여 학습시킨다.
5단계. 학습에 참여하지 않은 노드를 제거한다.
6단계. 미리 설정한 조건을 만족시키지 못하면 3단계로 돌아가서 반복한다.

3.3 앙상블 분류기

여러 가지 특징선택 방법들과 분류기들이 존재하지만 그중 완벽에 가까운 것을 찾기는 힘들다. 또한 선택되는 특징의 개수와 분류기의 매개변수 설정도 매우 어려운 문제이다. 분류기들은 환경에 민감하게 반응하기 때문에 환경을 어떻게 설정하는지 결정하는 것도 또한 매우 어렵다. 이러한 상황에서 한 가지 특징선택 방법이나 분류기만을 정해서 사용한다면 그것이 항상 좋은 결과를 내리라고는 기대할 수 없다. 분류기 앙상블은 이런 상황에서, 마치 사회적으로 어려운 문제에 대하여 위원회를 구성하고 결정하는 것처럼 유용하게 사용된다. 즉 분류기간의 결합을 통하여 더 좋은 성능을 기대할 수 있게 된다.

하나의 특징-분류기가 낼 수 있는 결과는 제한되어 있다. 따라서 다양한 결과를 얻을 수가 없다. 본 논문에서는 7가지의 특징선택법과 6가지의 분류기를 통한 총 42개의 특징-분류기 결과를 얻었으며, 이는 최대 42가지의 결과만 내는 것을 의미한다. 하지만 이렇게 나온 42개의 특징-분류기 중 임의의 것들을 선택하여 결합하면 2^{42}가지의 앙상블 결과를 얻을 수 있다. 이를 통하여 탐색할 공간이 단일 특징-분류기에 비하여 비교할 수 없이 넓어져 그만큼 다양한 결과를 낼 수 있다.

본 논문에서 사용한 앙상블 방법은 투표 방법과 가중투표 방법이며, 이들은 각 특징-분류기의 출력결과를 가지고 앙상블에 참여한다. 이진분류의 경우에 대하여 각각은 표 3과 같이 구할 수 있다. 가중투표 방법의 경우 가중치로는 그 특징-분류기의 인식률을 사용하였다.

표 3 앙상블 방법

(x가 입력인 경우, $e_i(x)$는 각 특징-분류기이며, $c_{1i}(x)$는 $e_i(x)$=1이면 1이고 그렇지 않으면 0, $c_{0i}(x)$는 $e_i(x)$=0이면 1이고, 그렇지 않으면 0이다. w_i는 i번째 특징-분류기의 인식률이다)

앙상블 방법	출력	조건
투표 결합	1	$\sum_i (c_{1i}(x)) > \sum_i (c_{0i}(x))$
	0	otherwise
가중투표 결합	1	$\sum_i (c_{1i}(x)w_i) > \sum_i (c_{0i}(x)w_i)$
	0	otherwise

3.4 GA를 이용한 앙상블 탐색

최적의 앙상블을 찾기 위하여 모든 앙상블에 대한 결과를 구하려면, $2^{42} \simeq 4\times10^{12}$ 가지에 대한 결과를 구해야 한다. 이를 모두 구해보는 것은 시간적으로 매우 비효율적이며, 이 상태에서 특징선택 방법이나 분류기가 추가된다면 모든 앙상블 수는 특징-분류기의 수에 지수

적으로 비례하여 증가하기 때문에 더욱 많은 시간이 소요된다. 만약 하나의 분류기가 추가된다면 7가지 특징선택 방법과 연결되어 7가지의 특징-분류기가 추가되기 때문에, 모든 앙상블 결과를 구하기 위해서는 $2^7=128$ 배의 시간을 추가로 소비해야 한다. 따라서 효과적으로 해공간을 탐색하는 기법이 필요하며, 본 논문에서는 GA를 이용하여 해 공간을 탐색하는 방법을 시도하였다. 본 논문에서 사용한 GA를 위한 염색체의 설계는 그림 4와 같다.

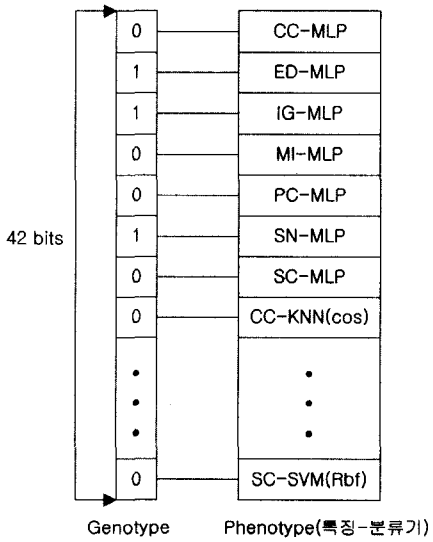

그림 4 GA의 염색체 구조

그림 4와 같이 염색체는 42비트로 이루어져 있다. 각 비트의 의미는 그 위치에 해당하는 특징-분류기를 앙상블에 사용할지 안할지의 여부이다. 각 비트는 비트의 순서에 대응하는 특징-분류기가 첫 번째 비트는 CC-MLP, 두 번째 비트는 ED-MLP, 마지막 비트는 SC-SVM(Rbf)인 식으로 사전에 정의되어 있다. 그림 4와 같은 경우는 2번째, 3번째, 6번째 비트가 1이고, 나머지 비트는 모두 0이므로, 주어진 염색체는 2번째, 3번째, 6번째에 대응하는 특징-분류기인 ED-MLP, IG-MLP, SN-MLP의 세 가지를 이용한 앙상블이라는 의미를 갖는다.

초기 집단이 임의로 생성된 후, GA에서 가장 먼저 수행되는 작업은 각 염색체의 적합도를 평가하는 것이다. 적합도 평가는 앙상블 결과를 대상으로 행하여지며, 본 논문에서는 앙상블 결과의 인식률을 적합도로 정하였다. 염색체가 그림 4와 같을 때의 투표 결합을 이용한 앙상블 결과와 적합도는 표 4와 같다.

표 4 앙상블 결과의 예

	결 과	인식률
2번째 특징-분류기	0 0 0 0 0 0 0 0 1 1 1 0 1 1	78.6%
3번째 특징-분류기	0 1 0 1 0 0 0 1 1 1 1 1 0 1	71.4%
6번째 특징-분류기	1 0 0 1 1 0 1 0 1 0 0 1 0 1	64.3%
투표 결합 앙상블	0 0 0 1 0 0 0 0 1 1 1 1 0 1	85.7%
실제 클래스	0 0 0 1 1 0 0 0 1 1 1 0 0 1	

표 4의 예는 결합 결과가 앙상블에 사용된 특징-분류기들의 인식률보다 향상된 경우로 85.7%가 나왔다. 각 염색체는 앙상블 결과에 대하여 이러한 방식으로 적합도를 평가받는다. 적합도를 평가받은 후 염색체들은 선택과정으로 들어가게 되는데, 본 논문에서는 각 염색체에 대하여 적합도에 비례하는 선택확률을 부여하였고, 선택방법으로는 룰렛 휠 규칙을 사용하였다. 선택된 염색체들은 메이팅 풀에 들어가서 두 개씩 짝지어진다. 짝지어진 염색체들은 교차 확률에 의하여 교차 여부를 결정하고, 교차가 결정되면 임의의 교차점을 정해서 유전정보를 교환한다. 그 후 염색체의 각 비트는 돌연변이 확률에 의하여 돌연변이 여부가 결정되는데, 돌연변이가 결정되면 그 비트는 0은 1로, 1은 0으로 바뀐다. 그림 5는 1번 염색체와 3번 염색체에 교차연산을 적용한 예이다. 이들은 여섯 번째 비트와 일곱 번째 비트 사이를 교차점으로 정해서 유전정보를 교환하였다.

그림 6은 교차연산을 마친 1번 염색체의 세 번째 비트에 돌연변이가 일어나 1에서 0으로 바뀌는 모습을 보여준다. 그림의 의미는 3번째 특징-분류기를 앙상블에 이용하지 않겠다는 것으로, 만약 0에서 1로 바뀐 비트가 있다면 이는 사용하지 않던 특징-분류기를 앙상블에 사용하겠다는 것이다. 이처럼 돌연변이는 원래의 정보가

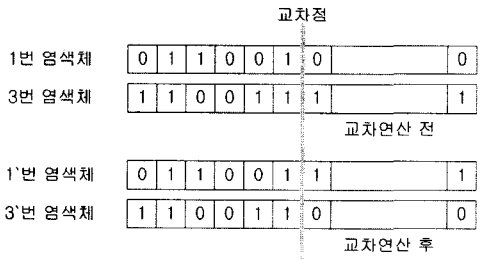

그림 5 교차연산

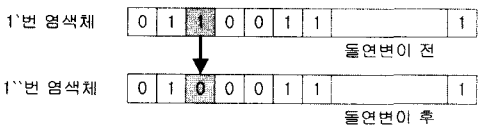

그림 6 돌연변이 연산

사라지거나, 새로운 정보가 들어온다는 것을 의미한다.

짝지어진 염색체들의 유전연산이 끝나면, 다음 세대의 부모가 되어 적합도를 평가하는 부분부터 다시 반복하게 된다.

4. 실험 및 결과

4.1 실험 데이타

실험 데이타로는 웹상에 공개되어 있는 유전발현 데이타인 림프종 데이타와, 대장암 데이타를 사용하였다.

(1) 림프종 데이타(Lymphoma dataset)

림프종 데이타(http://llmpp.nih.gov/lymphoma/)는 4026개의 유전자로 구성되어 있으며 총 47개의 샘플이 사용되었다. 이중 24개는 GC B-like DLBCL이고, 23개는 activated B-Like DLBCL이다[20]. 이 중 22개를 training에 이용하였고, 24개를 validation에 이용하였으며, 남은 1개를 test에 이용하였다. 과적합(overfitting)을 방지하기 위하여, 매번 다른 test data에 대하여 이러한 실험을 47회 반복한 leave-one-out cross validation을 수행하였다.

(2) 대장암 데이타(Colon dataset)

대장암데이타(http://www.sph.uth.tmc.edu:8052/hgc/default.asp)는 2000개의 유전자로 이루어져 있으며 62개의 샘플이 사용되었다. 이중 40개는 암 조직이고, 22개는 정상 조직이다[21]. 본 논문에서는 31개의 training data와 30개의 validation data 및 1개의 test data로 나눈 후, 과적합을 방지하기 위하여 62회 반복한 leave-one-out cross validation을 수행하였다.

4.2 실험 환경

실험은 크게 특징 선택, 분류, GA를 이용한 앙상블 탐색으로 이루어져 있다. 특징선택의 경우는 7가지 방법에 순위를 매긴 후 상위 25개의 유전자들을 선택하였다. 분류기의 경우, 인공신경망은 목표 인식률 98%, 최대 반복 회수 500회, 은닉노드 8개, 학습률은 0.01~0.5사이, 모멘텀은 0.9로 설정한 후 수십 회의 실험을 하였다. *K*NN의 경우 *k*값의 범위를 1~8로 주었으며, 유사도 측정 도구로는 피어슨 계수와 코사인 계수를 사용하였다. SASOM의 경우 4×4 크기의 지도를 초기에 사용하였다. SVM의 경우는 선형 함수와 RBF커널을 이용한 함수를 사용하였고, RBF의 경우 0.1에서 0.5사이의 감마값을 사용하였다.

예비실험 결과, GA를 이용한 앙상블 탐색에서 세대를 이루는 염색체의 수가 수십 개 이하일 때는 탐색할 수 있는 공간이 제한되어 있기 때문에 최적의 앙상블을 찾기가 어려웠다. 따라서 100개 이상인 100개, 200개, 500개, 1000개, 1500개, 2000개에 대하여 실험하였다. 염색체 선택 방법으로는 룰렛 휠 규칙과 순위 기반 방법을 사용하였으며, 교차율은 0.3부터 0.9까지 0.2 간격으로 정하였고, 돌연변이율은 0.01과 0.05를 사용하였다. 돌연변이에서 최적해로의 수렴을 빨리 시키기 위해서, 0에서 1로 바뀌는 비율을 1에서 0으로 바뀌는 비율의 반으로 정하였다. 실험의 종료 조건은 최적의 앙상블을 찾을 때까지 GA가 진화를 하되, 10만회를 넘길 경우에는 멈추게 하였다. 적합도는 validation data에 대한 분류 인식률로 정하였다.

4.3 단일 특징-분류기 실험 결과

두 가지 데이타에 대하여 각각 7가지 특징 선택법과 6가지 분류기의 조합에 대한 평균 분류 인식률은 표 5와 6과 같다. 림프종 데이타의 평균 인식률은 73.9%로 70.2%인 대장암 데이타보다 약간 높았고, 최대 인식률도 85.2%로, 대장암데이타의 최고 81.5%보다 상대적으로 높았다.

림프종 데이타의 경우 분류기로는 MLP가 우수한 결과를 보여주었고 그 뒤를 KNN이 이었다. 나머지 분류기의 성능은 70% 미만으로 좋지 못하였다. 특징선택법은 정보이득이 평균 80%이상의 가장 우수한 인식률을 보여주는 가운데 나머지 방법들은 큰 차이를 보이지 않았다.

한편 대장암 데이타의 경우도 MLP와 KNN이 다른 분류기에 비하여 우수한 성능을 보여주었고, 특징선택법은 전체적으로 비슷한 가운데 코사인 계수가 약간 우위를 보였다. 대장암 데이타에서는 60% 미만의 성능을 보이는 특징-분류기(SN-SASOM)도 존재하였다.

표 5 림프종 데이타에 대한 단일 특징-분류기 결과

	MLP	SASOM	SVM(L)	SVM(R)	KNN(C)	KNN(P)	평균
PC	77.6	67.6	66.4	66.8	78.4	78.0	72.5
SC	78.8	67.2	68.0	68.0	78.4	76.8	72.9
ED	75.2	62.8	66.4	66.4	76.0	77.6	70.7
CC	80.0	64.4	72.4	72.4	78.0	78.4	74.3
IG	85.2	75.2	77.6	77.6	81.6	83.2	80.1
MI	80.0	67.6	67.2	67.2	76.4	77.2	72.6
SN	81.2	70.8	68.0	68.4	78.8	79.2	74.4
평균	79.7	67.9	69.4	69.5	78.2	78.6	73.9

표 6 대장암 데이타에 대한 단일 특징-분류기 결과

	MLP	SASOM	SVM(L)	SVM(R)	KNN(C)	KNN(P)	평균
PC	78.2	67.7	64.5	64.5	69.4	76.6	70.2
SC	75.8	60.5	64.5	64.5	71.8	71.8	68.2
ED	75.8	64.5	65.3	65.3	76.6	77.4	70.8
CC	81.5	72.6	64.5	64.5	76.6	77.4	72.9
IG	77.4	63.7	66.1	66.1	72.6	75.0	70.2
MI	78.2	63.7	66.1	66.1	72.6	75.0	70.3
SN	74.2	58.9	64.5	64.5	77.4	74.2	69.0
평균	77.3	64.5	65.1	65.1	73.8	75.3	70.2

4.4 GA를 이용한 최적 앙상블 탐색 결과

최적의 앙상블을 찾기 전에 GA가 제대로 진화하고 있는지의 여부를 알아보기 위해서 평균 적합도의 추이를 살펴보았다. 이는 최적 앙상블의 발견이 해에 다가가는 과정으로 나온 것이 아니라 우연히 일어난 현상일수도 있기 때문이다. 그림 7은 림프종 데이타에서 투표결합을 하였을 때의 평균적합도 추이를 보여주는데, 전체적으로 세대가 지날수록 상승한다는 것을 알 수 있다. 전체적으로 100세대에 이르기까지 평균 적합도는 조금씩 상승하는 경향을 보여주었으며, 그 뒤로는 진동하며 수렴하는 모습을 보여주었다.

한편 집단을 이루는 염색체 수와 최적해를 발견하는 세대수와의 관계를 알아보기 위하여 집단을 이루는 염색체 수를 100개, 200개, 500개, 1000개, 1500개, 2000개로 변화시켜가며 실험을 해 본 결과, 그림 8과 같이 염색체 수가 증가할수록 평균적으로 더 빠른 세대에서 최고의 적합도를 보이는 염색체를 발견하였다. 그리고 돌연변이의 경우는 그 비율이 0.01일 때보다 0.05일 때 조금 더 빨리 최적의 앙상블을 찾아주는 경향을 보여주었다.

림프종 데이타와 대장암 데이타 모두 단일 특징-분류기의 평균 인식률은 약 70%로 그다지 높지 않았다. 따라서 상대적으로 앙상블을 통하여 결과가 많이 개선되는 것을 기대할 수 있는데, 실제로 두 가지 데이타에 대하여 GA는 단일 특징-분류기의 최고 결과보다 좋은 성능을 보이는 앙상블을 찾아주었다.

림프종 데이타의 경우는 validation data에 대하여 투표와 가중투표의 방법에서 모두 100%의 인식률을 보이는 앙상블을 찾아주었다. 이 데이타의 단일 특징-분류기 쌍의 결과는 62.8~85.2%에 퍼져있고, 평균적으로도 73.9%로 그다지 좋은 편이 아니다. 하지만 GA를 이용한 탐색은 상보적인 결합을 통하여 91.7~100%의 성능을 보이는 앙상블들을 발견하였다. 표 7은 두 가지 앙상블 방법에 의하여 100%의 앙상블 결과를 보이는 특징-분류기 쌍 중 GA가 발견한 것의 한 예를 보여준다. SC-SASOM은 겨우 62.5%의 인식률을 보여주지만, 투표 결합에서 다른 특징-분류기 쌍들과 상보적 결합을 통하여 우수한 앙상블을 구성하는 것에 제 역할을 다하고 있으며, CC-SASOM은 54.3%의 좋지 않은 인식률로 다른 특징-분류기 쌍들과 가중투표 결합을 통하여 우수한 앙상블의 구성원이 되는 것을 알 수 있다.

또한 대장암 데이타의 경우도 단일 특징-분류기 쌍의 결과 중 최고인 81.5%보다 우수한 90~96.7%의 결과를

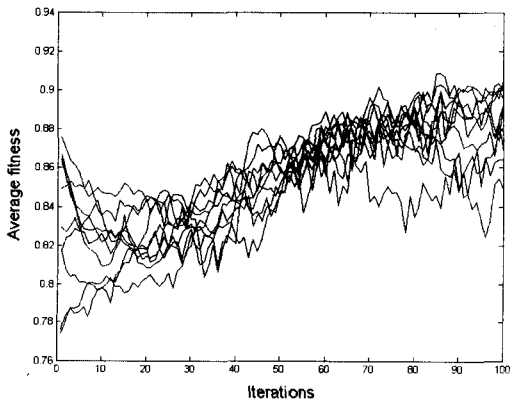

그림 7 림프종 데이타의 세대에 따른 평균 적합도

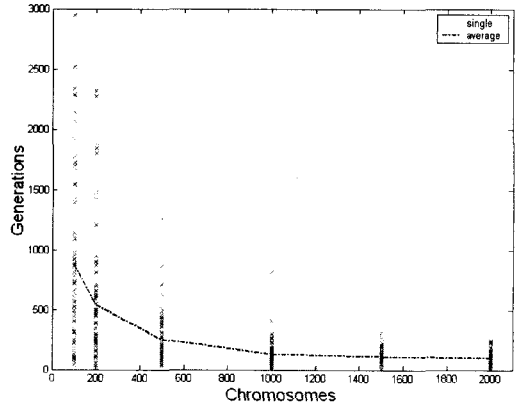

그림 8 염색체 수와 최적 앙상블 발견 세대와의 관계

표 7 림프종 데이타의 최적 앙상블

결합 방법	특징-분류기	인식률(%)
투표	CC-KNN(P)	75.0
	MI-KNN(P)	83.3
	SN-KNN(C)	79.2
	SC-SASOM	62.5
	IG-SVM(L)	91.7
	ensemble	100.0
가중투표	IG-KNN(C)	91.7
	MI-KNN(C)	83.3
	SN-KNN(C)	79.2
	SN-KNN(P)	79.2
	CC-SASOM	54.3
	IG-SASOM	83.3
	PC-SVM(R)	62.5
	ensemble	100.0

표 8 대장암 데이타의 최적 앙상블

결합 방법	특징-분류기	인식률(%)
투표	CC_MLP	86.7
	MI-MLP	70.0
	PC-MLP	73.3
	ED-KNN(C)	86.7
	ED-KNN(P)	86.7
	PC-KNN(P)	80.0
	SN-KNN(C)	76.7
	PC-SASOM	76.7
	ensemble	96.7
가중투표	MI-MLP	73.3
	PC-MLP	73.3
	CC-KNN(P)	83.3
	PC-KNN(C)	73.3
	SC-KNN(P)	70.0
	MI-SASOM	70.0
	ensemble	96.7

보이는 앙상블을 투표와 가중투표 방법 모두에서 발견할 수 있었다. 표 8은 대장암 데이타에서 최고의 인식률을 보이는 앙상블의 한 예를 나타낸 것이다. 투표 결합에 참가하는 단일 특징-분류기 쌍들은 70~86.7%의 인식률을 보여주지만 상보적 결합에 의하여 96.7%의 앙상블 인식률을 보여주고 있다. 가중투표 결합에 참가하는 단일 특징-분류기 쌍들도 평균적으로 73.9%의 인식률을 기록하였지만, 앙상블 결과 96.7%의 매우 뛰어난 인식률을 보여주었다. 따라서 앙상블을 통하여 상보적인 단일 특징-분류기 쌍의 조합을 찾아내었다고 할 수 있다.

한편, 실험 결과의 과적합 방지를 위한 leave-one-out cross validation결과, 림프종 데이타에서는 93.6%, 대장암 데이타의 경우는 87.1%의 성능으로, 단일 특징-분류기 쌍의 최고 인식률보다 높았다.

GA는 우수한 앙상블을 찾아내는 장점도 있지만, 더 유용한 점은 그 효율성에 있다. GA를 이용하지 않고, 42개의 특징-분류기 중 7개를 앙상블에 이용한 모든 결합에 대하여, 즉 $_{42}C_7 \simeq 2.8 \times 10^7$가지에 대하여 결과를 구하는 데에 걸린 시간을 측정한 결과 약 1시간 정도의 시간이 필요하였다. 따라서 2^{42}가지나 되는 모든 앙상블을 시도하는 것은 거의 불가능에 가까우며, 만약 이 상황에서 특징선택 방법과 분류기가 각각 하나씩 늘어나는 경우, 전체 앙상블을 구하기 위해서는 추정 결과 약 24만 년이 걸린다. 그러나 실제 GA를 이용하여 한 세대를 2000개의 염색체로 구성하고 룰렛 휠 선택 방법을 사용하여 500세대의 진화를 반복하였을 때, 선택 방법에 따라 약간의 차이가 있지만 30분이 걸리기 전에 연산이 끝났다. 실제로 한 세대를 2000개의 염색체로 구성하였을 때 평균적으로 약 100세대 안에 최적 앙상블을 찾아내었고, 집단을 100개의 염색체로 구성했을 때에도 평균적으로 900세대 안에 최적 앙상블을 찾아주었으므로, GA를 이용하는 것이 시간적으로 매우 효율적임을 알 수 있다. 이와 같은 실험결과들은 투표 결합과 가중투표 결합에서 모두 비슷한 양상을 보여주었다.

5. 결론 및 향후연구

표 9 모든 앙상블 탐색과 GA 탐색의 소요시간 비교

탐색 방법	구하는 앙상블 수	소요시간	비고
모든 앙상블 탐색	3×10^7	1시간	$_{42}C_7$
	4×10^{12}	15년(추정)	42개 특징-분류기
	6.4×10^{16}	240000년(추정)	56개 특징-분류기
GA 탐색	1×10^5	15초	룰렛 휠 선택
	1×10^6	5분	룰렛 휠 선택
	1×10^6	30분	순위 기반 선택

본 논문은 DNA 유전발현 데이타의 효과적인 분석을 위하여 생물의 진화과정을 모델로 한 방법인 GA를 사용하였다. 실험 결과 모든 데이타에 대하여 GA는 단일 특징-분류기 쌍의 최고 성능보다 우수한 앙상블을 매우 빠르게 찾아주었고, 평균 적합도의 추이를 통하여 우연에 의한 발견이 아니라 GA의 연산에 의한 것이라는 것을 알 수 있었다. 특히 림프종 데이타에 대해서는 100% 인식률을 보이는 조합을 찾아주었고, 대장암 데이타에 대해서도 96.7%의 높은 성능을 보이는 앙상블을 찾아주었다. 이와 함께 leave-one-out cross validation결과도 단일 특징-분류기 쌍의 인식률보다 훨씬 우수하였기 때문에, 제안한 방법의 우수함을 입증할 수 있었다. 비록 GA를 이용하지 않더라도 모든 앙상블을 비교하여 최적의 앙상블을 발견할 수 있겠지만, 이는 시간적으로 매우 비효율적이다. 그리고 특징-분류기의 수가 증가한다면 모든 앙상블을 구하기 위한 연산 시간은 그 수에 지수적으로 비례하여 증가하기 때문에, GA와 같은 효율적 탐색방법이 더욱 필요하다고 할 수 있다.

한편, 이번에는 룰렛 휠 선택 방법과 순위 기반 선택 방법 및 일점 교차를 사용한 일반적인 GA와 간단한 앙상블 방법만을 사용하였지만, 향후에는 더욱 다양한 방법을 시도해 볼 예정이고, 특징선택 방법과 분류기의 개수를 증가시켜서 실험을 할 예정이다. 염색체 설계에서도 단순하게 개별 특징-분류기의 앙상블 참여 여부만을 나타냈는데, 가중치까지 포함한 설계를 해 볼 예정이다. 데이타의 특성상 많은 수의 샘플을 사용하지 못하여 leave-one-out cross validation을 수행하였는데, 차후에 더 큰 규모의 데이타에 대하여 제안한 방법을 다시 한번 평가할 예정이다. 또한 GA가 찾은 최적의 앙상블에 대하여 그것에 속한 유전자들을 분석하지 못하였는데, 이는 생물학적으로도 분석할 가치가 있을 것이고 학제간의 연구주제로 매우 좋을 것이다.

참 고 문 헌

[1] T. R. Golub, et al., "Molecular classification of cancer class discovery and class prediction by gene-expression monitoring," *Science*, vol. 286, no. 15, pp. 531-537, October 1999.

[2] L. J. v. Veer, et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 31, pp. 530-536, January 2002.

[3] Y. H. Yang, et al., "Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, no. 4, e15, pp 1-10, 2002.

[4] 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, June 2001.

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

[6] M. P. S. Brown, et al., "Support vector machine classification of microarray gene expression data," *USCS-CRL-99-09*, pp. 1-23, June 1999.

[7] S.-B. Cho, and J.-W. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. of the IEEE*, vol. 90, no. 11, pp. 1744-1753, 2002.

[8] J. Quackenbush, "Computational analysis of microarray data," *Nature Reviews Genetics*, vol. 2, pp. 418-427, June 2001.

[9] T. M. Mitchell, *Machine Learning*, Carnegie Mellon University, 1997.

[10] S. Fuhrman, et al., "The application of Shannon entropy in the identification of putative drug targets," *BioSystems*, vol. 55. pp. 5-14, 2000.

[11] D. Thieffry, et al., "Qualitative analysis of gene networks," *Pacific Symposium on Biocomputing*, vol. 3, pp. 66-76, 1998.

[12] D. V. Nguyen, et al., "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39-50, 2002.

[13] S. Dudoit, et al., "Comparison of discrimination methods for the classification of tumors using gene expression data," *Technical Report 576*, Department of Statistics, University of California, Berkeley, 2000.

[14] Y. Xu, et al., "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer," *Cancer Research*, vol. 62, pp. 3493-3497, 2002.

[15] A. Ben-Dor, et, al., "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-584, 2000.

[16] R. P. Lippmann, "Pattern classification using neural networks," *IEEE Communications Magazine*, pp. 47-64, November, 1989.

[17] R. O. Duda, et al., *Pattern Classification*, 2nd Ed., Wiley Interscience, 2001.

[18] T. S. Furey, et al., "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906-914, 2000.

[19] H.-D. Kim and S.-B. Cho, "Genetic optimization of structure-adaptive self-organizing map for efficient classification," *Proc. of International Conference on Soft Computing*, pp. 34-39, October 2000.

[20] A. A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, February 2000.

[21] U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*. vol. 96, pp. 6745-6750, June 1999.

박 찬 호

2002년 2월 연세대학교 컴퓨터과학과(학사). 2002년 3월~현재 연세대학교 컴퓨터과학과 석사과정 재학중. 관심분야는 인공지능, 바이오인포매틱스, 패턴인식

조 성 배

정보과학회논문지 : 소프트웨어 및 응용 제 30 권 제 1 호 참조