산술 연산자 기반 유전자 프로그래밍을 이용한 암 분류 규칙 발견

(Rule Discovery for Cancer Classification using Genetic Programming based on Arithmetic Operators)

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

요 약 최근 생물정보 기술이 암 진단의 새로운 방법으로 관심을 모으고 있다. 다양한 기계학습 기법이 적용되어 우수한 결과를 얻고 있지만 의학 분야에서는 정확률이 높은 분류기뿐만 아니라 획득된 분류 규칙을 사람이 분석하고 이해할 수 있어야 한다. 생물정보 기술에서 많이 이용되는 유전자 발현 데이타는 데이타 내에 수천 내지 수만의 변수가 존재하며, 직접 이들 사이의 복잡한 관계를 표현하고 이해하는 것은 매우 어렵다. 본 논문에서는 이러한 어려움을 극복하기 위해 유전자 발현 데이타에서 분류에 유용한 특징들을 추출하고 산술 연산자 기반 유전자 프로그래밍으로 암 분류규칙을 생성하는 방법을 제안한다. 림프종 유전자 발현 데이타에 대하여 실험하여 96.6%의 인식률을 얻었으며, 획득된 분류 규칙을 분석하여 다양한 지식을 발견할 수 있었다.

키워드 : 유전자 프로그래밍, 지식발견, 암 분류, 특징추출

Abstract As a new approach to the diagnosis of cancers, bioinformatics attracts great interest these days. Machine learning techniques have produced valuable results, but the field of medicine requires not only highly accurate classifiers but also the effective analysis and interpretation of them. Since gene expression data in bioinformatics consist of tens of thousands of features, it is nearly impossible to represent their relations directly. In this paper, we propose a method composed of a feature selection method and genetic programming. Rank-based feature selection is adopted to select useful features and genetic programming based arithmetic operators is used to generate classification rules with features selected. Experimental results on Lymphoma cancer dataset, in which the proposed method obtained 96.6% test accuracy as well as useful classification rules, have shown the validity of the proposed method.

Key words : genetic programming, knowledge discovery, cancer classification, feature extraction

1. 서 론

암에 대한 정확한 판단과 분류는 의학 분야에서 매우 중요하고도 어려운 문제이다[1]. 정확한 암의 분류는 그에 대한 적절한 치료법과 약품 사용을 가능하게 하여 질병을 치료하고 환자의 생명을 구하는 중요한 일이다. 수세기에 걸쳐 다양한 암 분류 기법이 개발되었지만 대부분 전통적인 형태적 징후 분석에 기반하고 있다. 이들은 진료기반의 방법이어서 사람의 실수나 잘못된 해석 등이 발생할 수 있으며, 다른 종류의 암임에도 불구하고 유사한 징후가 나타나는 경우가 있기 때문에 많은 오분류를 초래하기도 한다. 이러한 한계를 극복하기 위해서 최근에는 사람의 유전자 정보를 이용한 분류기법이 연구되어 우수한 결과를 얻고 있다[2,3].

사람의 유전자 정보는 최근 주목받는 DNA micro-array 기술로부터 수집되며, 이들 유전자 발현 정보는 생명체에 관한 대량의 유전정보를 포함한다[2]. 많은 경우 유전자 발현 정보는 여러 종류의 암을 분류하는 데 유용한 정보를 제공한다. 하지만 유전자 발현 정보의 원시형태는 단순한 숫자들의 나열이기 때문에, 직접적으로 의미를 해석하거나 암을 분류하는 규칙을 발견하기는 매우 어렵다. 따라서 이것을 효과적으로 분석하기 위해

이 연구는 과학기술부가 지원하는 뇌과학연구 프로그램에 의하여 지원받은 것임
† 학생회원 : 연세대학교 컴퓨터과학과
hjinh@sclab.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터산업공학부 교수
sbcho@cs.yonsei.ac.kr
논문접수 : 2003년 11월 27일
심사완료 : 2004년 5월 4일

수년전부터 많은 방법이 연구되고 있다.

유전자 발현 정보의 분석은 크게 유전자 식별, 유전자 조절 네트워크 설계, 분류, 군집화의 4분야로 나뉜다. 정보이론은 유전자 식별 문제에, 이진 네트워크, 베이지안 네트워크, 역공학 기법 등은 유전자 조절 네트워크 설계에 각각 적용되었고, k 최근접이웃, 결정트리, 신경망, SVM 등의 다양한 기계학습 방법이 유전자 발현 정보 분류문제에, 자기구성지도, 계층적 군집화 등이 군집화 문제에 도입되었다[3]. 표 1은 유전자 발현 데이타의 분류에 관련된 연구들을 정리한 것이다.

이와 같이 다양한 인공지능 기술이 암을 분류하기 위해 적용되어 우수한 분류 성능을 보이고 있지만 이들 대부분은 사람이 직접 해석하기 어렵고, 많은 변수를 고려해야 하는 문제에서는 우수한 성능을 얻기 힘들다. 유전자 발현 정보를 이용한 암 분류는 그 특징의 수가 수천 개에 이르기 때문에 쉽게 우수한 분류기를 생성하기 어려우며, 또 사람이 이해할 수 있는 분류 규칙이 발견되지 않으면 신뢰하기 어렵다[3,4]. 따라서 거대한 해영역을 효과적으로 탐색하고 이해할 수 있는 분류 규칙을 얻기 위한 휴리스틱이 주목받고 있다. 유전자 알고리즘을 이용하여 일차논리 규칙을 추출하거나 IF-THEN 규칙의 집합을 획득하는 시도가 있었으며[4], 분석과 분류에 유용한 분류 규칙을 얻기 위해 유전자 프로그래밍을 이용하기도 하였다. 유전자 알고리즘이나 유전자 프로그래밍과 같은 진화연산은 적자생존을 기초로 한 자연선택 방법으로 해집단을 발전시켜 우수한 해를 얻는 기법으로 거대한 해영역 탐색에 효과적이고 염색체를 자유롭게 설계하여 사람이 이해하기 쉬운 해구조를 얻어낼 수 있어 최근 많은 분야에서 적용되고 있다. 본 논문에서는 유전자 프로그래밍을 이용하여 고차원의 유전자 발현 정보로부터 우수한 성능을 획득하고 사람이 이해할 수 있는 분류 규칙을 생성하는 방법을 제안한다.

2. 배 경

2.1 DNA microarray

생물체는 기본적으로 수천 개의 유전자와 RNA 및 단백질이 복잡하게 결합되어 다양한 기능을 한다. 전통적인 분자생물학은 단일 유전자를 기반으로 분석되었기 때문에 매우 제한적이었다. 최근 개발된 DNA microarray 기술은 기존 기술의 한계를 극복하고 초미세 단위로 유전 정보를 획득하여 하나의 칩 상에서 전체 염색체의 발현양상을 관찰하도록 한다. 따라서 보다 복잡한 생물체의 현상을 관찰하고 분석할 수 있게 되었다. DNA microarray는 용액이 투과되지 않는 딱딱한 지지체 위에 고밀도 cDNA를 고정시켜 수천 개 이상의 DNA나 단백질을 일정간격으로 배열하여 붙이고 분석 대상 물질과 결합시켜 그 양상을 분석하는 칩이다. 배열상의 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 각기 다른 형광물질을 동일한 양으로 합성한다. 이것을 레이저 형광 스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현정도를 얻게 되는데, Cy5/Cy3의 비율에 밑이 2인 로그를 취한 값을 그 셀의 발현정보 값으로 얻는다[1,2].

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

유전자 발현 정보는 그림 1에서의 과정과 같이 획득된다. 각 직사각형의 판이 하나의 배열이며 배열 안에 격자 모양으로 셀이 박혀있다. 각 셀에는 대응되는 유전자들이 심어져있다. 형광물질이 처리된 test와 reference에 해당하는 유전물질들이 해당 셀에 부착되며 셀에 달

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	84.6~
	대장암			94.1~
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

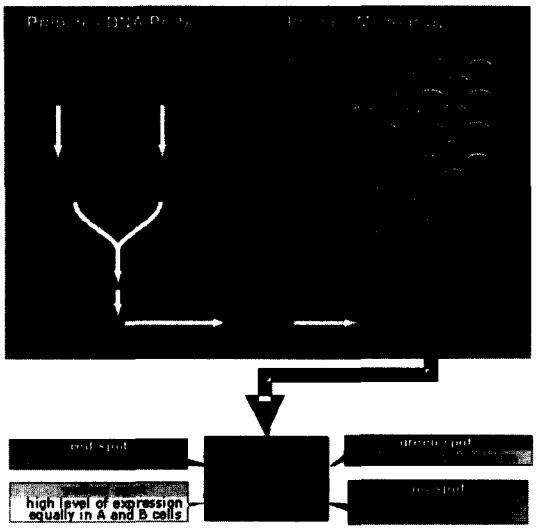


그림 1 DNA microarray 데이타 획득 과정

라붙은 정도에 따라서 다른 형광 정도를 보여준다.

하나의 microarray는 전체 염색체를 탐색하며 동시에 수천 개 유전자간의 상호관계를 분석할 수 있는 정보를 제공한다. 생물체에 관련된 다량의 유전정보를 얻기 위해서는 microarray의 이용이 필수적이다.

2.2 지식 발견

지식발견(Knowledge discovery)은 데이타로부터 자동적으로 지식을 추출하는 작업을 말한다. 추출되는 지식은 정확하고, 사용자가 이해할 수 있으며 흥미로운 것이어야 한다[4]. 지식발견이 적용되는 대표적인 문제로는 분류, 의존성 모델링, 군집화와 연관성 규칙 발견 등이 있으며, 결정 트리[5], 유전자 알고리즘[6,7], 유전자 프로그래밍 등의 방법이 많이 사용된다[8,9].

의료분야에서는 다량의 데이타로부터 유용한 지식을 발견하는 기술의 필요성이 증가하고 있다. 방대한 양의 데이타는 의학 전문가가 수작업으로 분석하기에는 거의 불가능하며, 아직 알려지지 않은 유용한 관계는 분석에 의해 쉽게 발견되지 않는다. 이러한 한계를 극복하기 위해 데이타로부터 유용한 정보를 분석하는 데이타마이닝이라 불리는 지식발견 기술이 의료정보분석에 사용되고 있다[4,8].

신경망 등의 기계학습 기법은 학습된 분류기로부터 사람이 이해할만한 분류규칙을 추출하거나 해석하기가 매우 어려우며, 특히 의학 분야에서는 학습된 분류기가 매우 높은 정확률을 가진다하더라도 쉽게 신뢰하지 못한다[10]. 기계학습 기법으로 얻어진 규칙이 전문가 등에게 해석이 가능하고 그 의미가 유효하다고 판명이 되어야 한다. 따라서 결정트리나 진화연산을 이용한 규칙 생성 방법이 보다 적합하며, 진화연산의 경우 염색체의 구조를 자유롭게 설계할 수 있기 때문에 보다 다양한 의미를 가지는 분류규칙을 생성할 수 있다[10-12].

2.3 유전자 프로그래밍

유전자 프로그래밍은 사용자가 명시적으로 프로그래밍을 하는 것이 아니라 컴퓨터로 하여금 주어진 문제를 해결하는 프로그램을 자동적으로 작성하도록 하기 위해 고안된 기술이다. 프로그램을 함수와 변수로 구성된 일종의 구조체로 간주하고 미리 정의된 문법에 어긋나지 않도록 이들을 구성한다[13]. 일반적으로 그림 2와 같이 root가 하나인 트리의 형태로 프로그램을 구성하며, 이것이 유전자 프로그래밍에서 개체의 유전자형이 된다.

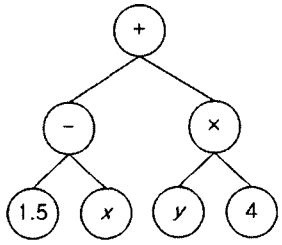


그림 2 유전자 프로그래밍의 유전자형

유전자 프로그래밍은 전통적인 유전자 알고리즘의 확장으로, 집단의 개체를 프로그램으로 정의하였다. 기본적인 동작과 특성은 유전자 알고리즘과 유사하지만, 개체의 표현형이 다르기 때문에 약간의 차이점이 있다. 유전자 프로그래밍의 해영역은 함수와 변수의 조합으로 발생할 수 있는 모든 가능한 프로그램이기 때문에 매우 광범위하다. 함수는 산술연산, 논리연산 및 사용자정의 연산 등 매우 다양하며, 이들 중 문제에 따라 적절히 선택하여 사용한다[10,11,14]. 이 기술은 특정한 문법을 가지는 프로그램으로 해결될 수 있는 모든 문제에 적용될 수 있으며, 문제영역에 의존적이지 않기 때문에 보다 다양한 분야에 사용될 수 있다. 최근에는 최적화문제나 어셈블리 코드의 진화, 진화하드웨어, 캐릭터행동진화 등의 문제에 많이 도입되고 있다[13].

다른 진화 연산 기법과 같이 유전자 프로그래밍은 다수의 개체로 구성된 집단을 기반으로 진화를 수행하며, 세대마다 기존 집단의 우수한 개체들에 유전 연산을 적용하여 새로운 개체를 획득한다. 우수한 유전 형질을 가지는 개체는 살아남고 그렇지 못한 개체는 도태되는 적자생존 원리를 바탕으로 최적의 해를 탐색한다. 그림 3~5는 유전자 프로그래밍에서 사용되는 유전 연산을 보여준다.

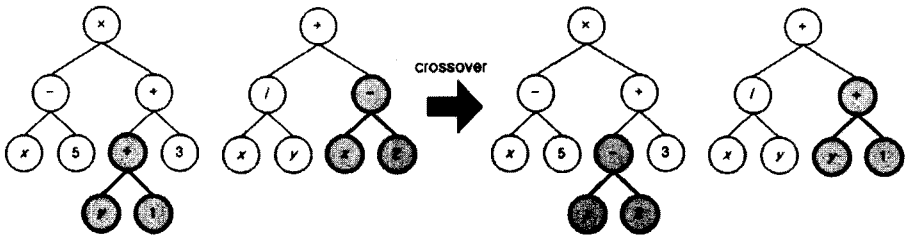


그림 3 유전자 프로그래밍의 교차 연산

3. GP 기반 분류 규칙 발견

본 논문에서는 그림 6에서와 같은 규칙발견 방법을 제안한다. 먼저 특징추출 단계를 거쳐 유용한 유전자만을 선택하여 유전자 발현 데이타의 특징 차원을 감소시키고, 선택된 유전자를 가지고 유전자 프로그래밍을 이용하여 우수한 분류규칙을 생성한다.

3.1 특징 추출

유전자 발현 정보의 유전자들이 모두 특정 질병과 연관되어 있는 것은 아니다. 따라서 특정 질병과 관련된 유전자를 선별하는 작업이 필요하며, 이를 특징선택 혹은 유전자선택이라 한다[3,15,16]. 특징선택은 학습속도를 향상시키고 잡음을 줄이는 효과가 있으며, 특징의 중요성을 측정하는 기준으로 순위를 매겨 선택하는 순위기반의 방법과 분류기와 연계된 학습데이타 자체의 특성을 이용하는 방법으로 구분된다. 순위기반 특징추출 방법은 단순하고 연산이 비교적 간단하지만 유전자 발현 데이타에 적용되어 우수한 성능을 보인다[15]. 따라서 본 논문에서는 상관계수, 유사도, 정보이론에 기반을 둔 총 7개의 순위-기반 특징추출 방법을 사용하였다. 각 선택 방법별로 유전자의 점수를 계산하고 상위에 랭크된 유전자를 선택하여 분류규칙 발견에 사용하였다.

(1) 상관계수 기반 방법

상관계수분석은 하나의 변수가 다른 변수와 관련성이 있는지, 또 관련 정도가 어느 정도인지를 분석하여 변수

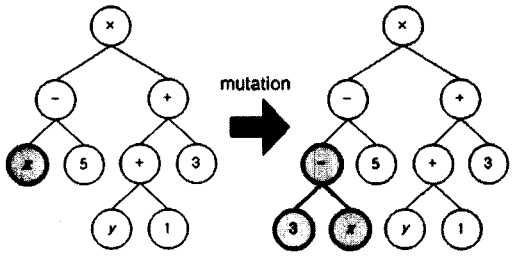


그림 4 유전자 프로그래밍의 돌연변이 연산

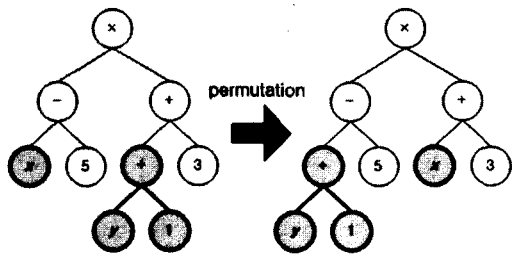


그림 5 유전자 프로그래밍의 치환 연산

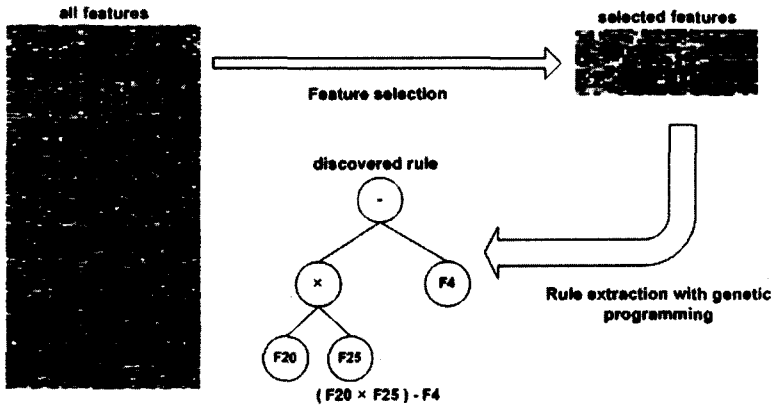


그림 6 제안하는 분류 규칙 발견 방법

간의 관련성을 파악하는 방법이다. 피어슨 상관계수는 상관계수분석에서 자주 이용되는 계수이며, 상관계수 r은 [-1, 1]의 값을 갖는다. 두 변수는 r의 값이 1에 가까울수록 양의 상관관계를 나타내어 서로 유사하며, r이 -1에 가까울수록 음의 상관관계가 되어 서로 반대로 작용한다. 또한 r이 0에 가깝다면 두 변수는 특별한 관계가 없음을 나타낸다. N개의 원소를 갖는 두 벡터 X와 Y사이의 피어슨 상관계수(PC)는 다음 식과 같이 정의된다.

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

한편 변수들의 값을 직접 이용하는 모수 분석과는 달리 양적 변수가 아닐 때 이용하는 스피어맨 상관계수와 같은 비모수 상관계수분석 방법이 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 다음 식과 같이 구할 수 있다.

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

(2) 유사도 기반 방법

상관계수분석이 두 변수의 상관정도를 분석한다면 유사도 측정법은 두 변수의 유사성을 측정한다. 두 변수간의 유사성을 거리로 나타내어 가까우면 유사성이 높음을 뜻한다. 유클리드 거리는 두 변수간의 기하학적 공간에서의 거리를 나타내며, 거리 값이 크게 나올수록 유사한 정도가 낮은 것이다. 두 벡터 X와 Y의 유클리드 거리(ED)는 다음 식과 같이 표현할 수 있다.

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

두 변수간의 유사성은 거리뿐만 아니라 두 변수사이의 각으로 나타낼 수도 있는데 각이 작을수록 같은 방향을 가리켜 서로 유사함을 나타낸다. 일반적으로 각을 직접 구하기보다는 그 각에 대한 코사인 값을 구하여 유사성을 측정한다. 두 변수사이의 각이 작을수록 코사인 값은 1에 가깝고, 변수가 완전히 반대의 방향을 가리킬 때 코사인 값은 -1로 나타나므로 코사인 계수는 [-1, 1]의 값을 가지며, 그 값이 1에 가까울수록 유사성이 높다. 두 변수 사이의 코사인 계수(CC)는 다음 식과 같이 나타낸다.

$$r_{\cos ine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

(3) 정보이론 기반 방법

전체 데이타로부터 의미있는 정보를 뽑아내는 척도로 정보이론에서 사용하는 정보이득(Information Gain), 상호정보(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 상호정보계수(MI)는 각각 다음 식과 같다.

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

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

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 다음 식과 같이 계산된다.

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

일반적으로 선택되는 특징의 개수가 너무 적으면 정보를 제대로 표현하지 못하는 문제가 있고, 너무 많으면 노이즈가 많이 포함된다는 문제가 있다. 따라서 적절한 특징 개수의 선택이 중요한데, 유전자 발현 데이타의 경우 20~70개 정도를 사용하였을 경우 큰 차이없이 비슷한 성능이 나왔다[3]. 따라서 본 논문에서는 모두 30개의 유전자들을 선택하여 분류규칙 발견에 사용하였다.

3.2 분류 규칙 발견

유전자 프로그래밍을 이용한 IF-THEN 규칙의 발견에는 다음과 같이 AND, OR 등의 논리연산과 (<, >, =) 등의 크기를 비교하는 연산자가 많이 사용되었다[4,12,17].

Rule 1 : IF ((*A*1 < 0.6) OR (*A*3 > 0.3)) THEN *class* 1
Rule 2: IF ((*A*2 = 0.7) AND (*A*1 > 0.7)) THEN *class* 2

이런 규칙은 비교적 해석이 쉬우나, 높은 정확률을 가지기 어려우며 변수들 사이의 복잡한 연관성을 표현하는데 한계가 있다. 반면에 변수들 사이의 보다 복잡한 연관성을 표현하기 위해서 함수 기반의 분류 규칙을 이용한 연구도 있었으나, 분류 규칙의 해석이 쉽지 않으며 성능에 있어서도 단순한 산술연산자로 구성된 분류 규칙보다 저조하기도 하였다[14].

본 논문에서는 보다 정확한 분류규칙을 발견하기 위

해 단순한 논리연산이 아닌 산술연산을 이용하여 분류 규칙을 구성한다. 특징추출 단계에서 뽑힌 30개의 특징과 기본적인 산술연산자(+, -, ×, /)를 이용하여 트리를 만들어 개체의 표현형으로 이용하였다. 표 2는 본 논문에서 사용된 산술연산자가 유전자 데이타에 대해 갖는 의미를 정의한 것으로, 이를 이용하여 발견된 분류 규칙을 해석한다.

표 2 산술 연산자 정의

산술 연산자	의미
+	• 클래스 1에 양의 영향 • 클래스 2에 음의 영향
-	• 클래스 1에 음의 영향 • 클래스 2에 양의 영향
×	배수 연관성
/	분배 연관성

분류규칙은 그림 7과 같이 구성되며, 대상 샘플의 특징값을 통해 계산된 eval()함수의 값이 부류를 결정한다. 따라서 계산된 값이 양수이면 부류 1, 음수이면 부류 2가 되도록 분류 규칙을 구성하였다.

IF eval($Individual_i$)≥0 THEN *class* 1
ELSE *class* 2

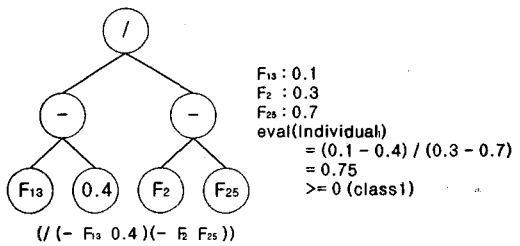


그림 7 개체 표현형과 분류규칙의 예

본 논문에서는 모든 산술연산자를 사용한 것이 아니라 표 3에서와 같이 3가지 형태를 사용하였다. 첫 번째 형식은 가장 복잡한 형태로 곱셈과 나눗셈 연산을 모두 사용한 경우이며, 두 번째와 세 번째 형식은 규칙을 보다 해석하기 쉽게 만들기 위해서 곱셈과 나눗셈 연산을 사용하지 않았다. 두 번째 형식의 경우에는 특징의 중요성을 측정하기 위해 각 특징에 가중치를 부여하여 설계하였다. 그림 8은 각 규칙의 형식의 예를 보여준다.

우수한 분류규칙을 발견하기 위해 기본적으로 학습 데이타에 대한 분류율을 유전자 프로그래밍의 적합도 함수로 사용한다. 또한 이해하기 쉬운 크기의 분류규칙을 얻기 위해 각 개체의 크기에 대한 평가를 적합도 평가에 추가한다. 일반적으로 동일한 성능을 내는 분류기의 경우 간단한 것이 일반화 능력이 뛰어나다고 알려져 있다.

$$fitness\ of\ individual_i = \frac{number\ of\ correct\ samples}{number\ of\ total\ train\ data} \times w_1 + simplicity \times w_2$$

$$simplicity = \frac{number\ of\ nodes}{number\ of\ maximum\ nodes}$$

w_1 : *weight* for *training rate*,
w_2 : *weight* for *simplicity*

4. 실험 및 결과

4.1 실험 환경

실험 데이타로는 웹상에 공개되어 있는 유전자 발현 데이타인 림프종 데이타를 사용하였다. 림프종 데이타(http://llmpp.nih.gov/lymphoma/)는 4026개의 유전자로 구성되어 있으며 총 47개의 샘플이 사용되었다[18]. 이중 24개는 GC B-like DLBCL이고, 23개는 activated B-Like DLBCL이다. 모든 샘플의 특징값은 정규화하여 사용하였다. 특징 수는 많지만 샘플 수가 매우 적기 때문에 하나의 샘플을 테스트 데이타로 놓고 나머지 샘플을 학습 데이타로 사용하는 것을 모든 샘플에 대해 반

표 3 분류 규칙의 형식

No.	+	-	×	/	Weighting	Complexity
1	Use	Use	Use	Use	Not-use	High
2	Use	Use	Not-use	Not-use	Use	Middle
3	Use	Use	Not-use	Not-use	Not-use	Low

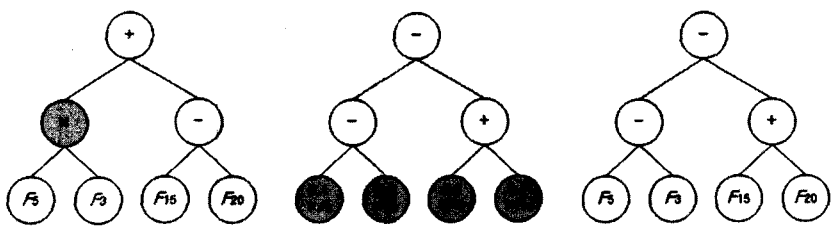


그림 8 분류 규칙의 형식 예

표 4 실험 파라메터

파라메터	값
집단 크기	100
최대 세대수	50000
선택율	0.6~0.8
교차율	0.6~0.8
돌연변이율	0.1~0.3
치환율	0.1
트리 최대 깊이	3
엘리트 유지전략	Yes

복하는 Leave-one-out cross validation으로 제안하는 방법을 평가하였다. 결과의 신뢰성을 위해 각 샘플에 대해 10회 반복하여 총 470(=10×47)회 실험의 평균을 최종 결과로 사용하였다.

실험은 먼저 7가지 특징선택 방법을 이용하여 상위 30개의 특징을 각각 추출하고 추출된 특징을 바탕으로 트리를 구성하였다. 유전자 프로그래밍의 설정은 표 4와 같다. 개체의 적합도 평가 함수에서 가중치 w_1과 w_2는 각각 0.9와 0.1로 설정하여 높은 성능의 분류율 획득을 우선적으로 중시하였으며 규칙의 크기도 어느 정도 고려하였다.

4.2 결과 분석

표 5는 각 특징선택 방법의 학습 데이타에 대한 인식률과 테스트 데이타에 대한 인식률을 보여준다. 학습 데이타에 대한 인식률은 정보이득을 제외한 대부분의 경우 100%에 가까웠다. 정보이득의 경우 테스트 데이타에 대한 인식률도 67.3%로 매우 저조하게 나왔으나 다른 특징추출 방법은 거의 90%이상의 인식률을 보였다. 특히 SN의 경우 테스트 데이타에 대해서 최고 96.6%의 인식률을 획득하였다. 특징선택 방법은 ED와 SN이 우수하였으며, 각 규칙 형식에 따른 결과에서는 세 번째 형식이 가장 좋은 성능을 거두었다. 그림 9는 각 규칙 형식과 특징선택 방법별로, 그림 10은 각 형식별로 성능을 그래프로 표현한 것이다.

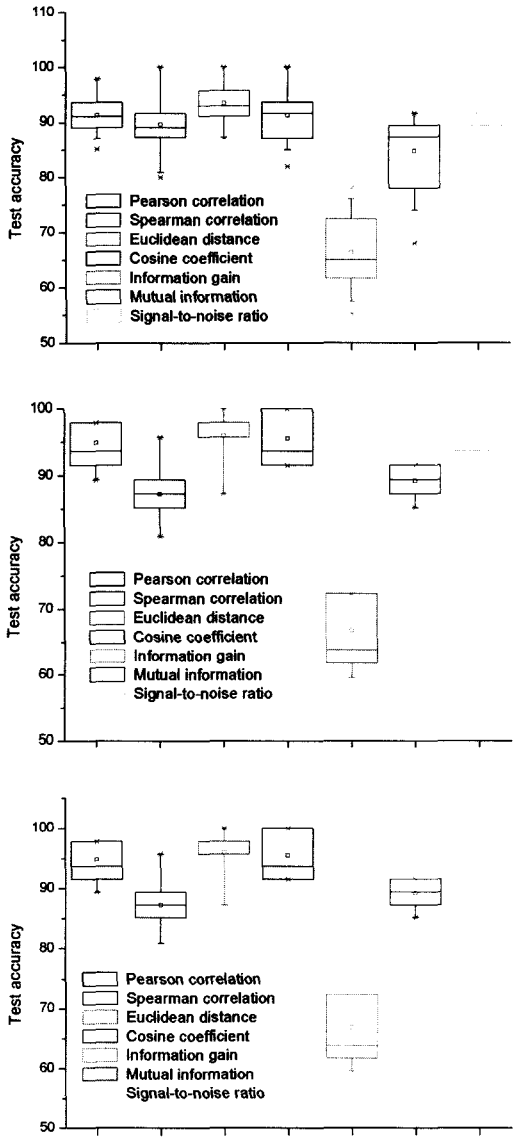


그림 9 규칙표현방법과 특징선택 방법에 따른 성능

표 5 인식률 결과 (1~3: 규칙의 형식)

특징 선택	학습 데이타			테스트 데이타			평균
	1	2	3	1	2	3	
PC	98.0	100	100	89.4	90.4	94.9	91.5
SC	96.8	100	100	88.2	93.4	91.3	90.9
ED	**99.0**	**100**	**100**	**91.8**	**92.8**	**95.9**	**93.5**
CC	98.9	100	100	86.9	92.3	95.5	91.5
IG	89.4	90.3	89.6	70.6	64.5	66.8	67.3
MI	96.6	99.7	100	77.4	87.4	89.1	84.6
SN	**98.7**	**100**	**100**	**89.9**	**93.6**	**96.6**	**93.3**
평균	96.6	98.5	98.5	84.8	87.7	90.1	

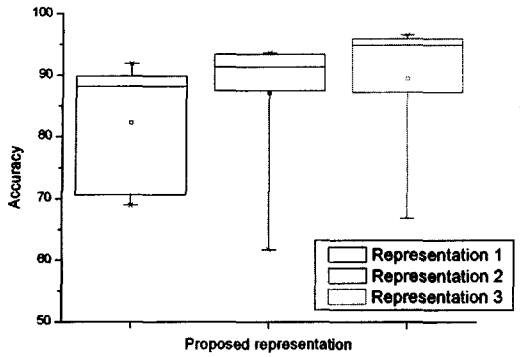
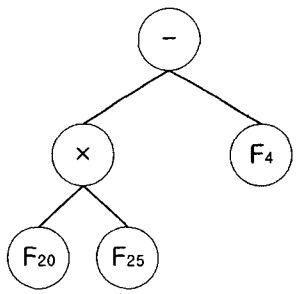


그림 10 규칙표현방법별 성능

그림 11은 첫 번째 규칙 형식으로 신호 대 잡음비 특징 중 단 3개의 유전자만을 이용하여 모든 샘플을 정확히 분류한 규칙이다. 이 규칙은 실험에서 가장 빈번히 발생한 규칙이며, 사용된 3개의 유전자에 대한 내용은 표 6과 같다. 사용된 유전자 중 하나는 이미 기능이 밝혀진 것이지만 나머지 둘은 아직 밝혀지지 않은 것으로 이에 대한 임상적 연구가 필요한 것이다.



$(F_{20} \times F_{25}) - F_4$

그림 11 첫 번째 형식을 이용한 100% 분류 규칙

그림 11의 분류 규칙은 앞서 설명한 산술 연산자의 정의를 이용하여 분석될 수 있다. 다음 수식에서와 같이 F4는 대상 샘플이 부류 2에 속하도록 영향을 주며, F20과 F25는 배수 결합을 하여 부류 1에 속하도록 하는 영향을 준다.

$$Class1 \Leftarrow -F_4 \qquad Class2 \Leftarrow +F_4$$
$$Class1 \Leftarrow +(F_{20} \times F_{25}) \qquad Class2 \Leftarrow -(F_{20} \times F_{25})$$

본 논문에서는 산술 연산자를 이용한 분류 규칙의 분류성능을 비교 평가하기 위해서 신경망을 이용한 실험을 추가로 수행하였다. 먼저 그림 11의 분류 규칙에서 사용된 3개의 유전자를 이용하여 신경망을 학습하여 분류를 수행하였고, 두 번째로 신호 대 잡음비 특징 30개를 모두 이용하여 신경망을 학습하여 분류 성능을 측정하였다. 신경망은 세 개의 층으로 구성되었으며 출력층은 2개의 노드로 구성하여 각 부류를 나타내도록 하였다. 은닉층은 2~10개의 노드로 구성하고, 학습률은 0.01~0.1, 모멘텀은 0.7~0.9로 다양하게 설정하였다. 제안하는 방법에 사용한 leave-one-out cross validation으로 성능을 평가하였다. 첫 번째 실험 결과로, 학습 데이타에 대한 인식률은 98%, 테스트 데이타에 대한 인식률은 97.8%이 되어 제안하는 방법에 미치지 못하는 결과를 획득하였다. 두 번째로 모든 특징을 사용한 경우에는, 학습 데이타에 대한 인식률과 테스트 데이타에 대한 인식률이 모두 95.7%로 처음 실험보다 떨어지는 것을 확인하였다. 이는 유전자 프로그래밍을 이용하여 분류 규칙을 생성하였을 때, 특징 추출 효과도 함께 얻었음을 말해준다.

그림 12와 13은 두 번째와 세 번째 규칙 표현형을 사용하여 얻은 100% 분류 규칙으로, 표 7과 표 8에 각각 사용된 유전자에 대한 정보가 기술되어 있다. 이들 분류 규칙도 신호 대 잡음비 특징들로 구성되어 있다.

각각의 분류 규칙에는 기존에 기능이 알려진 유전자들이 속해 있으며, 특별히 림프종 암과 관련이 깊은 #86와 #1914 유전자가 사용되었다[19,20]. 이는 발견된 분류 규칙이 기존의 연구에 비추어 타당성이 있으며, 이미

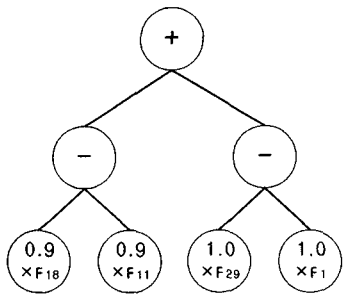


그림 12 두 번째 표현형을 이용한 100% 분류 규칙

표 6 사용된 특징

특징번호	DNA 번호	내용
F20	75	ets variant gene 6 (TEL oncogene); 14671
F25	2467	*core binding factor alpha1b subunit=CBF alpha1=PEBP2aA1 transcription factor =AML1 Proto-oncogene=translocated in acute myeloid leukemia; Clone=263251, 17823
F4	1277	Unknown UG Hs.136345 ESTs; Clone=746300, 19274

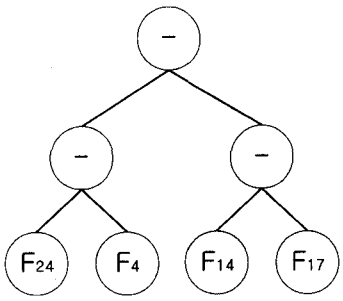


그림 13 세 번째 표현형을 이용한 100% 분류 규칙

알려진 유전자를 바탕으로 함께 사용된 유전자 기능의 연구에 도움을 줄 수 있음을 의미한다. 표 9는 각 분류 규칙에서의 특징이 클래스에 미치는 영향을 보여준다.

제안하는 방법과 기존의 분류규칙 생성방법의 성능비교를 위해서 본 논문에서는 결정트리를 이용하여 동일한 leave-one-out cross-validation 방식으로 각 특징 추출방법에 대해서 총 329(=47×7)번의 실험을 수행하였다. 사용한 결정트리 알고리즘은 전통적으로 많이 이용되는 C4.5[5]와 최근 C4.5를 개량하여 상업용으로 배포하고 있는 See5[21]를 이용하였다. 그림 14는 제안하는 방법과 결정트리와의 성능을 보여주며, 제안하는 방법이 평균 분류율과 최고 분류율에서 보다 우수함을 확인할 수 있었다. 분류규칙 분석에 있어서 제안하는 방법은 각

표 7 그림 12의 규칙에 사용된 특징 정보

특징번호	DNA 번호	내용
F18	1636	CXCR5=BLR1=B-cell homing chemokine receptor=L1; Clone=31, 4297
F11	1246	*FAK=focal adhesion kinase; Clone=795352, 17333
F29	86	*BCL-2; Clone=342181, 17646
F1	1268	*CD10=CALLA=Neprilysin=enkepalinase; Clone=200814, 15864

표 8 그림 13의 규칙에 사용된 특징 정보

특징번호	DNA 번호	내용
F24	684	Unknown; Clone=1352715, 14377
F4	1279	*Unknown; Clone=825199, 19288
F14	1914	Lymphotoxin-Beta=Tumor necrosis factor C; Clone=1320296, 13297
F17	680	*Unknown; Clone=1372162, 19541

표 9 발견된 분류 규칙 분석

대상	분석 과정	분석 결과
그림 12의 분류 규칙	$Class1 \Leftarrow +F_{18} : +(+F_{18})$ $Class1 \Leftarrow -F_{29} : +(-F_{29})$ $Class1 \Leftarrow +F_{11} : +(+F_{11})$ $Class1 \Leftarrow -F_{1} : +(-F_{1})$	• F18은 클래스 1에 양의 영향 • F29는 클래스 1에 음의 영향 • F11은 클래스 1에 양의 영향 • F1은 클래스 1에 음의 영향
그림 13의 분류 규칙	$Class1 \Leftarrow +F_{24} : +(+F_{24})$ $Class1 \Leftarrow -F_{4} : +(-F_{4})$ $Class1 \Leftarrow -F_{14} : -(+F_{14})$ $Class1 \Leftarrow +F_{17} : -(-F_{17})$	• F24는 클래스 1에 양의 영향 • F4는 클래스 1에 음의 영향 • F14는 클래스 1에 음의 영향 • F71은 클래스 1에 양의 영향

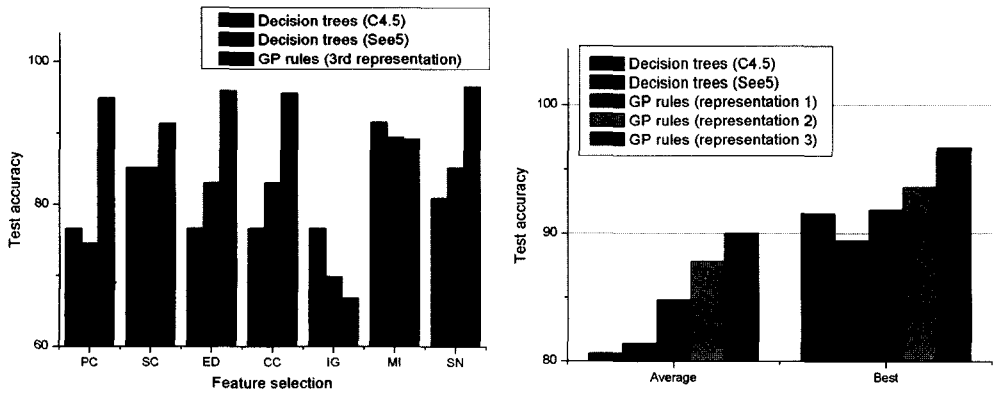


그림 14 결정 트리와의 성능 비교

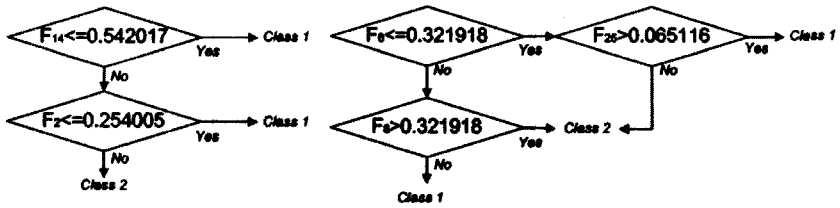


그림 15 결정트리 알고리즘으로부터 얻어진 분류 규칙

특징의 클래스에 대한 영향력을 측정한 반면, 결정트리는 전통적으로 많이 사용되는 논리적 분류규칙을 이용하였기 때문에 다른 측면의 정보를 얻는다. 그림 15는 결정트리 알고리즘을 통해 얻어진 규칙을 보여준다. 얻어진 규칙의 크기는 깊이 3의 트리 정도로 제안한 방법의 규칙과 거의 비슷한 수준이었다.

5. 결 론

본 논문은 DNA 유전자 발현 데이타의 효과적인 분석을 위하여 규칙발견 및 표현에 유용하고 생물의 진화과정을 모델로 한 유전자 프로그래밍을 사용하였다. 특징차원은 크고, 샘플의 수는 매우 적은 유전자 발현 데이타로부터 유의한 분류규칙을 추출하는 것은 어려운 문제이다. 특징추출을 수행하여 유용한 특징을 선택하고, 유전자 프로그래밍을 이용하여 선택된 특징들로 산술구조의 규칙을 생성하였다. 진화과정에서 얻어진 다양한 분류규칙으로부터 100%의 분류성능을 내는 산술구조의 분류규칙을 발견할 수 있었다.

유전자 프로그래밍은 사람이 이해할 수 있는 수준의 분류규칙 생성에 유용하다. 논리 및 산술구조를 복합적으로 사용한 분류규칙은 보다 높은 설명기능과 성능을 낼 것으로 예상되며, 단일 분류규칙만을 사용할 때보다 다양한 분류규칙을 생성하여 이들을 결합한다면 보다 높은 분류 성능을 얻을 수 있을 것이다. 제안한 방법은 2-클래스 분류를 가정하기 때문에 향후에는 계층적으로 분류 규칙을 구성하여 다중 클래스 분류가 가능하도록 확장하는 연구가 필요하다.

참 고 문 헌

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

[2] A. Brazma and J. Vilo, "Gene expression data analysis," *Federation of European Biochemical Societies Letters*, vol. 480, pp. 17-24, 2000.

[3] C. Park and S.-B. Cho, "Genetic search for optimal ensemble of feature-classifier pairs in DNA gene expression profiles," *Int. Joint Conf. on Neural Networks*, pp. 1702-1707, 2003.

[4] K. Tan, et al., "Evolutionary computing for knowledge discovery in medical diagnosis," *Artificial Intelligence in Medicine*, vol. 27, no. 2, pp. 129-154, 2003.

[5] J. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.

[6] D. Goldberg, *Genetic Algorithms in Search, Optimaization*, and Machine Learning, Addison-Wesley, 1989.

[7] K. DeJong, et al., "Using genetic algorithms for concept learning," *Machine Learning*, vol. 13, pp. 161-188, 1993.

[8] A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," *Advances in Evolutionary Computation*, pp. 819-845, 2002.

[9] C. Hsu and C. Knoblock, "Discovering robust knowledge from databases that change," *Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 69-95, 1998.

[10] C. Zhou, et al., "Discovery of classification rules by using gene expression programming," *Proc. of the 2002 Int. Conf. on Artificial Intelligence*, pp. 1355-1361, 2002.

[11] C. Bojarczuk, et al., "Discovering comprehensible classification rules using genetic programming: A case study in a medical domain," *Proc. of the Genetic and Evolutionary Computation Conf.*, pp. 953-958, 1999.

[12] I. Falco, et al., "Discovering interesting classification rules with genetic programming," *Applied Soft Computing*, vol. 1, no. 4, pp. 257-269, 2002.

[13] J. Koza, "Genetic programming," *Encyclopedia of Computer Science and Technology*, vol. 39, pp. 29-43, 1998.

[14] J. Kishore, et al., "Application of genetic programming for multicategory pattern classification," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 242-258, 2000.

[15] H.-H. Won and S.-B. Cho, "Neural network ensemble with negatively correlated features for cancer classification," *Lecture Notes in Computer Science*, vol. 2714, pp. 1143-1150, 2003.

[16] J. Bins and B. Draper, "Feature selection from huge feature sets," *Proc. Int. Conf. Computer Vision 2*, pp. 159-165, 2001.

[17] S. Augier, et al., "Learning first order logic rules with a genetic algorithm," *Proc. of the First Int. Conf. on Knowledge Discovery & Data Mining*, pp. 21-26, AAAI Press, 1995.
[18] A. Alizadeh, et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.
[19] O. Monni, et al., "BCL2 overexpression in diffuse large B-cell lymphoma," *Leuk Lymphoma*, vol. 34, no 1-2, pp. 45-52, 1999.
[20] P. Koni and R. Flavell, "A role for tumor necrosis factor receptor type 1 in gut-associated lymphoid tissue development: genetic evidence of synergism with lymphotoxin β," *J. of Experimental Medicine*, vol. 187, no. 12, pp. 1977-1983, 1998.
[21] Data mining tools See5, http://www.rulequest.com/see5-info.html

홍 진 혁
2002년 연세대학교 기계전자공학부 정보산업전공 졸업. 2002년~2004년 연세대학교 컴퓨터과학과 석사. 2004년~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 지능형 에이전트, 인공생명, 패턴인식, 진화게임

조 성 배
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 1 호 참조