영상의 대표값과 유전자 프로그래밍을 이용한 자궁경부세포진 영상 인식

김재륜* 김백섭** 이헌길*** 하진영***
강원대학교 컴퓨터정보통신공학과*
한림대학교 컴퓨터공학과**
jaeryun@mail.kangwon.ac.kr*
bskim@sun.hallym.ac.kr**
{hklee, jyha}@cc.kangwon.ac.kr***

Cervical Cell Classification using Genetic Programming and Central tendency of Image

Jae-Ryun Kim* Baek-Sub Kim** Heon-Kil Lee*** Jin-Young Ha***
Dept. of Computer Information Engineering, Kangwon Natioinal University* ***
Dept. of Computer Science, Hallym University**

요 약

유전자 프로그래밍은 프로그램 자동생성 도구이다. 문제를 해결하는 프로그램 코드를 프로그래머가 직접 구현하는 것이 아니라, 적절한 초기값만을 입력하여 컴퓨터가 스스로 적합한 해를 찾아내도록 하는 방법이다. 유전자 프로그래밍은 생물의 진화개념에서 얻어진 여러 아이디어를 사용하여 최적화된 해를 찾아낸다. 본 논문에서는 세포영상인식 문제를 해결하기 위하여 유전자 프로그래밍을 사용하였다. 실험에 사용된 영상은 자궁경부세포진 영상이다. 여러가지 종류와 상태의 세포들이 뒤섞여 있어 분석하기에 힘들다는 것이 이 영상의 특징이다. 주어진 문제는 샘플 영상이 암인가 아닌가를 판별하는 것이다. 유전자 프로그래밍을 적용하기 위하여 사용한 특징값들은 영상에서 찾을 수 있는 가장 단순한 대표값들과, 산술 및 논리연산자들이다. 실험결과 실제 인식기 제작에 바로 적용하기엔 무리가 있지만, 80%정도를 제대로 판별해 낼수 있었다. 인식률이 낮은 이유는 사용한 특징들이 영상의 정보를 잘 흡수하지 못했기 때문이라 여겨지고, 앞으로 지나치게 복잡하지 않으면서 영상의 특징을 잘 표현하는 특징값들을 찾는 것이 향후과제이다.

1. 서론

자궁경부세포진 검사는 여성의 자궁경부암 여부를 판별하는 널리 알려진 검사법이다. 전통적인 자동화 시스템은 먼저 카메라에서 얻어진 세포영상을 분석하여 세포핵, 세포질, 아티팩트로 분할 된 영상을 얻는다. 이 분할된 영상에서 세포에 대한 특징들을 추출한다. 이 추출된 특징에 병리기사의 지식을 적용하여 세포의 암여부를 판별한다. 그런데, 촬영된 영상은 여러 종류와 상태의 세포들이 뒤섞여 있고, 촬영 때의 환경에 따라 달라져 분석하기가 매우 어렵다. 이 영상을 분석하기 위해서 본 논문에서는 유전자 프로그래밍 방법을 사용하였다.

유전자 프로그래밍은 프로그램 자동생성 도구이다. 프로그램을 자동으로 생성한다는 것은, 문제를 해결하는 프로그램을 프로그래머가 직접 구현하는 것이 아니라, 적절한 초기값만을 입력하여 컴퓨터가 스스로 적합한 해를 찾아내도록 하는 것이다.

본 논문에서는 먼저 유전자 프로그래밍에 대해 간단히 소개한 뒤, 대표값을 이용하여 유전자 프로그래밍을 적용하는 과정을 보이고 결론을 맺는다.

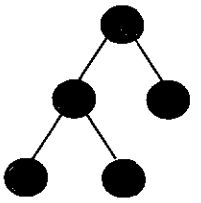
2. 유전자 프로그래밍

유전자 프로그래밍은 1992년 J.R Koza 에 의하여 소개 되었다[1]. 기존에 존재하던 유전자 알고리즘을 기계학습 분야에 맞도록 변형 발전시킨 것이 유전자 프로그래밍이다. 유전자 프로그래밍은 컴퓨터가 스스로

문제 해결을 위한 코드를 생성하도록 하기 위해서 생물의 진화에서 얻은 힌트를 기본 아이디어로 사용한다. 생물은 자신과 닮았지만 결코 똑같지 않은 자식을 거듭 생산함으로써 여러 세대가 교체된 후 조상보다 더 환경에 잘 적응한 개체를 탄생시킨다고 알려져 있다. 마찬가지로 유전자 프로그래밍은 거듭된 프로그램코드의 생성과 그에 대한 평가를 통하여 최종적으로 환경(문제)에 잘 적응한 프로그램을 생성해 낸다.

2.1 프로그램

유전자 프로그래밍에서의 프로그램은 보통 트리 구조로 표현된다. 이 트리 구조가 바로 프로그램이자 유전자 본체에 해당한다.



F(x) = (x*y)/2

<그림 1>

<그림 1>에 프로그램 트리의 예를 보인다. 이 예는 오직 단순한 산술연산 수식일 뿐이지만, 산술 연산뿐만 아니라 논리 연산 및 프로그램 함수 등 프로그램 트리로 표현할 수 있는 것은 무엇이든지 유전자 프로그래밍을 적용할 수 있다. 유전자 프로그래밍은 유전자에 해당하는 트리를 생성하고 평가하는 과정을 반복하여 문제의 해를 찾아낸다.

유전자 프로그래밍을 구성하는 요소는 terminal set, function set, fitness function, 종결조건이다.

2.2 Terminal Set

프로그램 트리에서 리프노드를 terminal 이라 부른다. Terminal은 프로그램에서 데이터에 해당하는 부분이다. 프로그램 코드의 데이터 후보들의 집합이 바로 terminal set이다.

2.3 Function Set

프로그램 트리에서 리프노드를 제외한 노드를 function이라 부른다. Function은 프로그램에서 함수에 해당하는 부분이다. 프로그램 코드의 연산과 함수에 해당하는 부분이다. 이들 연산과 함수의 후보들의 집합이 바로 function set이다.

2.4 Fitness Function

프로그램 트리의 적합도 기준이 되는 함수이다. 이 함수를 이용하여 프로그램 트리의 적합도를 산출한다.

2.5 종결 조건

유전자 프로그래밍의 수행을 끝내기위한 조건이다. 보통은 error function을 이용하여 error가 문턱값 이하로 떨어지면 수행을 종료한다.

2.6 유전 연산자

유전 연산자란 유전자 프로그래밍에서 프로그램 트리를 변형시키는 방법이다. 부모 세대의 프로그램들에 유전연산자를 적용하여 변형된 자식 세대를 얻어 낸다. 유전 연산자의 종류는 <표 1>에 정리되어 있다.

유전 연산자	설명
Reproduction	우수한 부모 프로그램을 그대로 자식세대로 복사함
Crossover	두 부모 프로그램의 일부를 서로 뒤바꿈
Mutation	부모 프로그램의 일부를 임의로 변형시킴
Permutation	부모 프로그램의 연산 순서를 뒤바꿈

<표 1> 유전 연산자

2.7 유전자 프로그래밍의 수행 순서

1. Terminal set과 function set을 이용하여 초기 세대 무작위로 생성
2. 집단의 각 프로그램을 평가하여 적합성 결정
3. 종결 조건이 만족되면 수행 종료
4. 유전 연산자를 이용하여 새로운 세대 생성
5. 2~4 계속 수행

3. 세포영상에의 적용

유전자 프로그래밍을 영상처리에 적용하기 위해서는 영상의 특징을 잘 대표할 수 있는 terminal set 과 function set을 찾아내는 것이 중요하다[2].

본 논문에서는 영상의 픽셀값들을 하나의 모집단으로 간주, 통계적인 대표값을 뽑아내어 유전자 프로그래밍의 terminal set으로 사용하였다. function set으로는 단순한 사칙연산과 논리연산을 사용하였다. 이러한 값들을 사용함으로 히스토그램을 분석한 것과 같은 효과를 얻으려 하였다. Terminal set과 function set,

건은 $f_1 < 0.01$ 이다. 한 세대의 개체수는 300, 최대 세대수는 500을 사용하였다.

Terminal set	평균, 중위수, 최빈수, 최대값, 최소값, 분산, 표준편차, 상수들(0,1,2,3, 255), 불린 상수(true, false)
Function set	Add, Sub, Div, Mul, GT, LT, MIN(), MAX(), IF -THEN-ELSE, AND, OR, XOR, NOT
Fitness function	f1 = (FP + FN)/A f2 = FN/F FP : 정상에 대하여 비정상 판별 개수 FN : 비정상에 대하여 정상 판별 개수 A : 총 데이터 수 F : FP+ FN F1과 f2를 비교하여 최소한 f에 대한 fn의 비율이 0.5보다 작게 나오도록 함

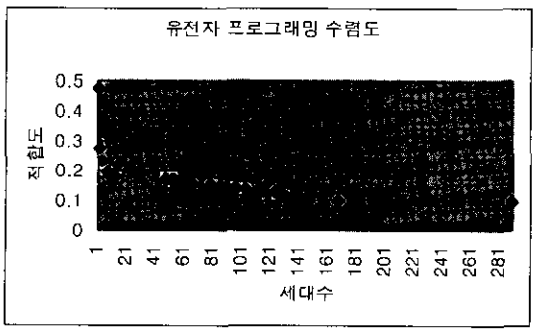
<표 2>

4. 실험 및 결과

작업한 환경은 PentiumII 400, 128M RAM, Win98, JAVA 1.2이다. 640 x 480의 영상을 80 x 60의 조각으로 나누어 실험 데이터를 준비하였다. Training set에는 정상 200, 비정상 200개를, test set에는 정상 100, 비정상 100개를 사용하였다. 프로그램은 100세대 이상 적합도에 변화가 없었기에 390세대에서 강제 종료 하였다. 총 수행시간은 13시간 29분 20초가 걸렸고 인식율은 <표 3>에 정리되어 있다.

	Training Set	Test Set
Error rate	0.0975	0.155
FP error rate	0.105	0.13
FN error rate	0.09	0.18

<표 3> 인식율



<그림 2>

fitness function이 <표 2>에 정리 되어 있다. 종결조

```
(Not (GT (Sub (Sub (Min (Min min (Mul max (Sub
standardDeviation mean))) 0) (Sub (Min (Sub min
mode) (Sub (Min (Min standardDeviation (Mul
max (Sub standardDeviation mean))) 0) (Sub
(Min (Sub min max) variance) max))) mode))
(Sub (Sub (Sub (Div (Sub (Sub (Div (Sub (Sub
(Sub (Sub (Div (Sub 0 variance) 0) max)
standardDeviation) variance) variance) 0) (Div
(Sub max (Min (Div (Sub variance variance) 0)
(Sub (Sub (Sub standardDeviation (Sub (Mul 2
max) (Div max min))) variance) 2))) 0)) variance)
(Min min (Min standardDeviation (Mul mode (Sub
min mean))))) standardDeviation)
standardDeviation) max)) variance))
```

<표 3> 결과 트리(복잡도94)

5. 결론 및 향후과제

자궁경부암 세포진 영상의 암여부를 판별하는 문제를 해결하기 위하여, 유전자 프로그래밍을 적용시켜 보았다. 히스토그램을 분석한 것과 동일한 효과를 얻기 위하여, 통계의 대표값들을 terminal set으로 사용하였고, 산술연산과 논리연산들을 function set 으로 사용하였다. 실험 결과 80%정도의 인식률을 보였다.

지나치게 복잡하지 않으면서 영상의 특징을 잘 표현할 수 있는 특징값들을 찾는 것이 향후과제이다.

6. 참고 문헌

[1]. Koza, John R., Genetic Programming : On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems), MIT Press, 1992.

[2]. Riccardo Poli, Genetic Programming for Image Analysis