

# 데이터 마이닝의 분류 규칙 발견을 위한 유전자알고리즘 학습방법

김대희, 박상호  
안동대학교 컴퓨터공학과

## Genetics-Based Machine Learning for Generating Classification Rule in Data Mining

Dae-Hee Kim, Sang-Ho Park  
Dept. of Computer Science, Andong Nat'l University  
E-mail : eogml@wail.co.kr, spark@andong.ac.kr

### 요 약

데이터(data)의 홍수와 정보의 빈곤이라는 환경에 처한 지금, 정보기술을 이용하여 데이터를 여과하고, 분석하며, 결과를 해석하는 자동화 된 데이터 분석 방안에 높은 관심을 가지게 되었으며, 데이터 마이닝(Data Mining)은 이러한 요구를 충족시키는 정보기술의 활용방법이다. 특히 데이터 마이닝(Data Mining)의 분류(Classification) 방법은 중요한 분야가 되고 있다. 분류 작업의 핵심은 어떻게 적당한 결정규칙(decision rule)을 정의하느냐에 달려 있는데 이를 위해 학습능력을 가지고 있는 알고리즘이 필요하다. 본 논문에서는 유전자 알고리즘(Genetic Algorithm)을 기반으로 하는 강력한 학습방법을 제시했으며, 이러한 학습을 통해 데이터 마이닝(Data Mining)의 분류시스템을 제안하였다.

### 1. 서론

컴퓨터 기술의 발전과 자료저장 구조에 대한 기술적인 발전으로 대용량의 데이터 처리가 가능해지면 기업들은 많은 양의 데이터를 축적할 수 있게 되었다. 그러나 축적된 정보를 어떻게 효과적으로 운영, 관리하여 기업이익에 필요한 정보 또는 지식을 가공해 낼 수 있는가의 여부가 중요한 관심사가 되었다. 데이터의 양이 오늘날과 같이 방대하지 않았던 과거에는 소수의 전문가들이 통계기법이나 질의를 통해 데이터를 분석하고 요약된 결과를 보고서 형식으로 제공해 주었다. 하지만 이러한 방법은 데이터의 양이 증가하고 요구하는 지식이나 정보가 다양해짐에 따라 효율성이 떨어지고, 분석을 통해 얻을 수 있는 정보의 품질도 기대하기 어렵다. 따라서 데이터의 홍수와 정보의 빈곤이라는 환경에 처한 지금, 정보기술을 이용하여 데이터를 여과하고, 분석하며, 결과를 해석하는 자동화

된 데이터 분석 방안에 높은 관심을 가지게 되었으며, 데이터 마이닝(Data Mining)은 바로 이와 같은 요구사항을 충족시키는 새로운 정보기술의 활용방법이다.

데이터 마이닝(Data Mining)은 대량의 데이터로부터 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술이라고 정의할 수 있다. 예를 들면, 전자상거래를 위한 웹 서버인 경우에는 소비자가 방문한 웹 페이지와 구매한 물건과 소비자의 특징을 보관하고 있기 때문에 이 데이터를 분석하면 각각의 사용자에 맞는 웹페이지를 동적으로 그때 그대 생성해 주거나, 웹 페이지의 성능을 높이고 수행속도를 빠르게 할 수 있다. 또한 모든 소비자에게 동일한 웹 페이지를 제공하는 것이 아니라 소비자의 관심에 따라 다른 웹 페이지를 동적으로 만들어 제공하는 개인화(personalization) 서비스를 가능하게 할 수도 있다.

데이터 마이닝은 <그림 1>에서 보여주듯이 여러

분야에서 활용 가능하며, 데이터를 유용한 정보로 변환할 수 있는 새로운 기술과 도구로 중요한 연구 대상으로 자리잡게 되었다.[1]



(그림 1) 데이터 마이닝의 활용분야

데이터 마이닝 방법은 데이터를 분석하여 어떤 종류의 정보를 찾고자 하는가에 따라 일반화/요약, (Generalization & summarization), 데이터 군집화 (clustering), 연관 (Association), 분류 (classification) 등 여러 가지가 있다.

데이터 일반화란 데이터베이스에서 많은 관련된 데이터를 낮은 개념 레벨에서 높은 개념 레벨로 추상화시키는 작업이다. 일반화/요약화 규칙은 데이터베이스 내의 사용자가 지정한 부분에 대해 일반적인 특성이나 요약된 고급 뷰를 제공한다. 통상 여러 추상화 레벨에 있는 데이터에 관하여 일반화된 뷰를 제공하는 것은 바람직하다.

클러스터링이란 레코드들을 유사한 특성을 지닌 몇 개의 소그룹으로 분할하는 작업을 뜻한다. 이 때 유사성 때문에 함께 모여진 개체의 셋을 클러스터라 한다. 클러스터링 작업은 먼저 필수 객체들이 셋으로 모여지고 이로부터 일련의 규칙이 유도된다.

연관 규칙은 레코드의 셋에 대하여 아이템의 집합 중에 존재하는 친화도나 패턴을 찾아내는 규칙이다.

분류는 데이터 마이닝에서 가장 많이 사용되는 작업의 하나로 주어진 데이터와 각각의 데이터에 대한 클래스가 주어진 경우, 그것을 이용해 각각의 클래스를 갖는 데이터들은 어떤 특징이 있는지 분류 모델을 만들고, 새로운 데이터가 있을 때 그 데이터가 어느 클래스에 속하는지 예측하는 것을 뜻한다. 특정한 분류 작업의 핵심은 어떻게 적당한 결정 규칙 (decision rule)을 정의하느냐에 달려있다.

데이터 마이닝은 학습능력을 가지고 있어야 한다. 학습능력이라는 것은 미리 정해진 순서와 절차에 따라 문제를 해결하는 전통적인 프로그래밍 방식과는 달리 주어진 레코드들로부터 추론이나 직관적 판단을 통해 스스로 문제 해결을 피하는 능력을 말한다.

본 논문에서는 데이터 마이닝 분야의 분류 규칙 발견을 위해 유전자 알고리즘을 기반으로 학습방법을 제시하고자 한다.

## 2. 데이터 마이닝 분류규칙과 유전자 알고리즘

### 2-1. 데이터 마이닝 분류규칙

분류 (Classification)는 마이닝 분야에 있어서 주요 연구분야중 하나로 그 목적은 과거에 알고 있는 데이터베이스 정보로부터 새로운 데이터베이스 정보를 분류해낼 수 있는 분류 규칙을 생성해 내는 것이다. 따라서 분류 문제는 다음과 같이 기술되어 진다.

“트레이닝 셋이라 불리는 입력 데이터가 여러 레코드로 구성되어지고 각 레코드는 애트리뷰트를 갖는다. 각 레코드는 클래스(또는 그룹) 레이블로서 표현되어 진다. 분류의 목적은 데이터에 나타난 특징 (feature)을 사용해서 입력 데이터를 분석해서 각 클래스에 대한 정확한 모델을 개발하는 것이다. 클래스 값은 앞으로 클래스 레이블을 모르는 테스트 데이터 (test data)를 분류하는데 쓰인다.”

여기서 사용되는 입력 데이터는 복잡성을 가지며, 끊임없이 변화하는 새로움에 직면하게 된다. 이러한 상황에 적합한 것이 학습이다. 학습은 과거 경험에 대한 일반화에 의하여 자신의 성능을 향상시킨다. 학습에 의한 경험은 끊임없는 새로움과 복잡성 속에서 관련 있는 규칙들이 존재하기만 하면 미래의 행동을 좌우할 수 있다.

학습시스템은 다음과 같은 몇몇 문제들과 접하게 된다.

1. 반복적인 데이터와 잘못된 혹은 관련 없는 자료를 가진 문제.
2. 행동에 대한 계속적이고 빈번한 실시간의 요청을 하는 문제.
3. 암시적인 또는 부정확하게 정의된 목표들을 가진 문제.
4. 긴 행동의 연속을 요구하는 빈약한 보수 혹은 강화를 가진 문제.

이와 같은 문제들을 다루기 위해서는 학습 시스템은 다음과 같은 것을 해야 한다.

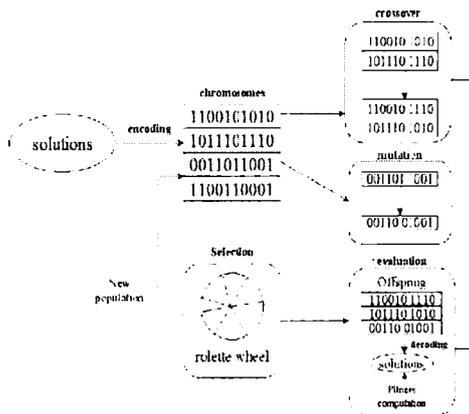
1. 그 환경에서 목표와 관련된 규칙성을 나타나게 하는 범주들을 창안.
2. 환경이 모형을 꾸준히 개선하기 위해 목표로 가는 동안 만나는 정보의 흐름을 이용.
3. 목표로 가는 동안에 만나는 단계 설정 범주에 알맞은 행동들을 할당.

이러한 학습시스템에는 기호적인 방법(symbolic), 강화기반 방법(reinforcement-based), 그리고 유전자 기반 방법(genetic-based)등이 있다.

### 2-2. 유전자 알고리즘

유전자 알고리즘은 유전자의 진화를 모방한 경험적 탐색법의 하나로서 최근 여러 분야에서 많이 활용되고 있다.[2] 유전자 알고리즘은 탐색공간이 크거나 분석적으로 해를 찾을 수 없는 문제에 대해 해결책을 제시할 수 있다. 유전자 알고리즘은 선택적 도태나 돌연변이 같은 생물 진화의 원리로부터 착안된 알고리즘으로서 확률적 탐색이나 학습 그리고 최적화를 위한 한 가지 기법이다. 유전자 알고리즘을 다음과 같이 설명할 수 있다.

첫째, 문제에 맞는 염색체(chromosome)의 구조를 정의하고 그 유전자의 적합도를 다지는 적합도 함수를 설계한다. 또 문제에 맞는 유전자 조작을 정의한다. 둘째, 초기 모집단을 생성한다. 이때 모집단의 숫자도 제한한다. 셋째, 모집단을 가지고 유전자 조작을 한다. 넷째, 조작된 유전자들의 염색체를 적합도 함수를 써서 평가한다. 다섯째, 설정된 확률대로 점수가 높은 유전자(gene)들의 모집단에서의 비율을 높이고 낮은 점수들은 없앤다. 여섯째, 확률에 의해 점수가 높은 순서대로 복제(reproduction)과정을 통해 더욱 많은 자가복제를 하고 그것들끼리의 교배(crossover)를 통해 새로운 모집단을 생성한다.



(그림 2) 유전자 알고리즘의 구조

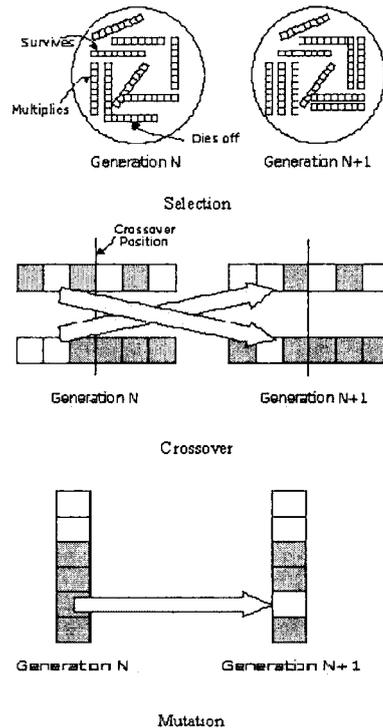
이러한 과정을 반복하면서 높은 점수를 가지는 유전자들을 계속 확장시켜 나가는 것이다. 이러한 과정

을 <그림 2>로 보이고 있다. 하지만 이 방법대로 진행하면 어떤 경우 국지 최적해(local minima)에 빠질 위험이 있으므로 돌연변이(mutation)를 만들어 임의의 영역을 탐색한다. 교배를 통해 원하는 해를 수렴하고 또한 변이를 통해 임의의 영역으로 탐색공간을 넓혀 적합도를 계산한다. 이러한 과정을 반복함으로써 원하는 해에 수렴하고자 하는 것이다.

유전자알고리즘에서 선택(Selection)은 개체집단의 크기를 유지하면서 환경에 대한 적합도(Fitness)를 평가하여 복제, 유지, 소멸의 과정을 거쳐 세대를 거듭할수록 적합도를 향상 시켜 나간다.

교배연산자는 확률적으로 결정되는 Crossover Position을 기준으로 하여 개체간의 유전자를 교환하고 결합하여 부모의 유전자 특성을 가지는 새로운 자식 개체를 생산하는 과정을 말한다.

돌연변이는 유전자를 일정한 확률로 변화시켜 집단 내에 전혀 새로운 개체를 생성하게 된다.



(그림 3) 유전자 알고리즘의 연산자

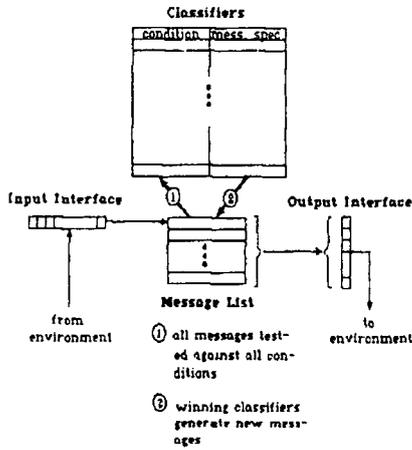
### 3-3. 분류시스템(Classifier System)

분류시스템은 구문론으로 단순한 문자열 규칙을 학습해 가는 기계학습시스템이다. 분류시스템은 모든 규

칙들이 동일한 단순 형태를 갖는 병렬적이고, 메시지 전달방식의 규칙 기반 시스템이다. 분류시스템은 세가지의 주요한 구성요소로 이루어져 있다.

1. Rule and message system
2. Apportionment of credit system :  
The bucket brigade
3. Genetic Algorithm

Rule and message system은 <그림 4>에 나타나 있듯이 입력인터페이스(input interface), 출력인터페이스(output interface), 메시지 리스트, Classifiers로 구성되어 있다.



(그림 4) Rule and Message System

기본적인 수행 주기는 다음과 같다.

- 단계 1. 입력 인터페이스로부터의 모든 메시지들을 메시지 리스트에 더한다.
- 단계 2. 메시지 리스트의 모든 메시지들과 모든 분류자들의 모든 조건들을 비교하여 모든 짝(조건들을 만족하는)을 기록한다.
- 단계 3. 몇 개의 분류자의 조건 부분을 만족하는 모든 짝들의 집합 각각에 대하여, 행동 부분에 의하여 명시된 메시지를 새로운 메시지의 리스트에 붙인다.
- 단계 4. 메시지 리스트에 있는 모든 메시지들을 새로운 메시지들의 리스트로 대체한다.
- 단계 5. 메시지리스트에 있는 메시지들을 출력인터페이스로 변환하고, 그것에 의하여 시스템의 현재 출력을 산출한다.
- 단계 6. 단계 1로 돌아간다.

이 시스템에서는 다음과 같이 몇 가지를 정의한다.

메시지 <message> := {0,1}  
 분류자 <classifier> := <condition>:<message>  
 조건 <condition> := {0,1,#}

Classifier store에 <표 1>과 같이 네 개의 Classifier가 저장되어 있다고 가정하자.

메시지 리스트에 0111이 저장되면, 이 메시지는 1번 classifier에 만족하고 새로운 메시지 0000을 기록한다. 그 다음에 메시지 0000은 2번, 4번 classifier에 만족하고 새로운 메시지에 1100과 0001을 기록한다. 메시지 1100은 3번, 4번 classifier에 만족하고 1000기록하고 메시지 1000은 4번 classifier에 만족하고 그 처리는 끝난다.

Index	Classifier
1	01##:0000
2	00#0:1100
3	11##:1000
4	##00:0001

(표 1) Four Classifiers

Apportion of credit system은 분류시스템들에 대한 점수 할당 문제를 해결하기 위해 고안되었다. 표2를 통해 Apportion of credit system을 이해할 수 있다.

표2의 예제는 표 1과 같은 four classifiers를 가지고, 각 분류자는 강도(Strength)를 200을 할당시키고 메시지 0111을 입력받는다. Bid 계수는 0.1로 한다. Apportion of credit system은 시스템에 대한 분류자의 전체적인 유용성을 반영하기 위해 Bid 계수에 의해 강도를 조절한다. 각 단계(t=0~t=5)에서는 각 만족된 분류자는 그것의 강도에 기반하여 입찰하고, 단지 가장 높은 값을 가진 분류자들만이 다음 단계에 메시지 리스트에 그것들의 메시지들을 놓는다. 이를 반복하여 마지막 단계에서는 시스템에 Payoff를 준다.

Index	Classifier	t=0				t=1			
		Strength	Messages	Match	Bid	Strength	Messages	Match	Bid
1	01##:0000	200		E	20	180	0000		
2	00#0:1100	200				200		1	20
3	11##:1000	200				200			
4	##00:0001	200				200		1	20
Environment		0	0111			0	0111		

Index	Classifier	t=4				Final(t=5)	
		Strength	Messages	Match	Bid	Strength	Payoff
1	01##:0000	220		E	20	220	
2	00#0:1100	208				208	
3	11##:1000	196				196	
4	##00:0001	156	0001			206	50
Environment		20				20	

(표 2) A Simple Classifier System by Hand-Matching and Payments

Genetic system은 분류 시스템의 규칙 발견에 사용한다. 유전자 알고리즘은 높은 강도의 분류자들을 “부모들”로서 선택하여, 부모 분류자들로부터 구성 성분을 재 조합함에 의하여 “자식”을 만들어 낸다. 그 자식은 조건이 만족될 때 활성화되고 검사되면서, 시스템에서 약한 분류자들을 대체하고 경쟁을 유도한다.

### 3. 실험 및 결과

#### 3-1. 문제영역

실험은 4개의 특징과 각 특징은 4개의 가능한 값으로 구성된 인공적으로 만들어진 간단한 문제에서 실행했다. 9개의 목적 개념(target concept)을 구성하고 목적개념의 복잡도를 규칙당의 규칙의 수와 관련된 특징의 수를 증가시킴으로서 변동했다. 선언의 수는 1부터 4까지이며, 결합의 수는 1부터 3까지 이다. 각 목적개념은 nDmC로 표시되는데, 여기서 n은 선언의 수이고 m은 결합의 수이다.[3]

예를 들어 F1, F2, F3, F4로 표시된 4개의 특성(feature)이 있고, 각 특성은 {v1, v2, v3, v4}의 4개의 값을 가진다.

모든 목적 개념은 다음의 형태를 가진다.

4DmC == d1 ∨ d2 ∨ d3 ∨ d4  
 3DmC == d1 ∨ d2 ∨ d3  
 2DmC == d1 ∨ d2  
 1DmC == d1

nD3C 목적 개념을 위하여, 다음의 형태를 가진다.

d1 == (F1=v1) & (F2 = v1) & (F3 = v1)  
 d2 == (F1=v2) & (F2 = v2) & (F3 = v2)  
 d3 == (F1=v3) & (F2 = v3) & (F3 = v3)  
 d4 == (F1=v4) & (F2 = v4) & (F3 = v4)

nD2C 목적 개념을 위하여, 다음의 형태를 가진다.

d1 == (F1=v1) & (F2 = v1)  
 d2 == (F1=v2) & (F2 = v2)  
 d3 == (F1=v3) & (F2 = v3)  
 d4 == (F1=v4) & (F2 = v4)

nD1C 목적 개념을 위하여, 다음의 형태를 가진다.

d1 == (F1=v1)  
 d2 == (F1=v2)  
 d3 == (F1=v3)  
 d4 == (F1=v4)

목적 개념들의 각각을 위하여, 집합에 있는 모든 경우들은 긍정(1), 혹은 부정(0)으로 표시한다. 이 실험에서는 긍정과 부정의 집합을 가진 문제에서 잘 수행

하는 규칙을 위해 규칙 집합의 공간을 탐색하는 분류시스템(Classifier System)이다.

<그림 5>는 Classifier System의 알고리즘이다.

detectos()는 environmental to classifier detectors이며, timekeep()은 time coordination routines이며, match\_classifier()는 rule and message system이며, aoc.scs()는 apportionment of credit routine이다. ga.scs()는 유전자 알고리즘(genetic algorithm) 부분이다.

```

/* Classifier system */
main()
{
    initialization();
    detectors();
    for(i=0;i<=timekeep();++i)
    {
        timekeeper();
        environment();
        detectors();
        match_classifiers(message);
        aoc();
        effector();
        reinforcement();
        if(timekeeper())
        {
            ga();
        }
    }
}

```

(그림 5) Classifier System

#### 3-2. 결과

위에서 언급한 시스템의 성능을 제안한 문제영역에서 평가했다. 실험에 사용한 파라미터 값으로 교배율은 1로 정했으며, 돌연변이율은 0.02로 정했다. 집단의 크기는 실행 시간상 이 실험에서는 1000개의 집단크기를 정했다.

많은 문제로부터 그것들의 일반적인 경향과 규칙을 알아내고, 새로운 문제에서 어떤 결과물을 예측하는 데에는 많은 방법들이 제시되고 있지만 이 논문에서는 Rule and message system, Apportionment of credit system, Genetic Algorithm을 기반으로 한 Classifier System을 사용하였다. 앞에서 제시한 nDmC 문제에서 규칙을 올바르게 생성했으며 수행 능력도 우수하다.

### 4. 결론

데이터 마이닝은 대량의 데이터로부터 데이터에 함

축되어있는 지식이나 규칙을 찾아내는 방법이다. 데이터 마이닝은 지식이나 규칙을 찾아내는데 있어서 질의에 대한 정교한 데이터를 찾아주거나, 기업에 전략적 판단을 위한 분석 정보를 추출하고, 의사 결정을 지원하거나, 현재의 데이터에서 미래에 대한 예측정보를 추출할 수도 있어야 한다. 이 논문에는 유전자알고리즘을 기반으로 하는 강건한 학습방법을 제시했으며, 이러한 학습을 통해 데이터 마이닝의 핵심기능인 분류 시스템을 제안하였다. 제안된 시스템은 점수 할당(bucket brigade algorithm)과 규칙 발견(genetic algorithm)을 통해 학습하는 매우 병렬적이고, 메시지 전달 방식의 규칙 기반의 시스템이다.

향후 연구로는 제안된 시스템의 특성을 고려하여 수행능력을 향상시킬 수 있는 새로운 유전자 조작이나 방법에 대한 연구가 필요하며, 보다 다양한 데이터를 가지고 실제 응용에 적용하여 본 논문에서 제안한 방법의 성능을 측정하는 연구이다.

#### [참고문헌]

- [1] U.M.Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R.Uthurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press, 1996.
- [2] M. Gen and R. W. Cheng, "Genetic Algorithms and Engineering Design", John Wiley and Sons, New York, 1997.
- [3] Kenneth A.de Jong, William M.spears,and Diana F.Gordon, "Using Genetic Algorithms for Concept Learning, Machine Learning", 13, pp.161-188, 1993.
- [4] J.G Carbonell, "Machine Learning: Paradigms and Methods", The MIT Press, 1990.
- [5] R. Agrawal, G. Psaila "Active Data Mining", Proc. of the 1st Int'l Conference on Knowledge Discovery and Data Mining, Montreal, August 1995.