유전자 알고리즘을 이용한 데이터 마이닝의 분류 시스템에 관한 연구

Using Genetic Rule-Based Classifier System for Data Mining

한 명 묵*
Myung-Mook, Han

요 약

데이터 마이닝은 방대한 데이터 자료로부터 숨어있는 지식이나 유용한 정보를 추출하는 과정이다. 이러한 데이터 마이닝 알고리즘은 통계학, 전자계산학, 그리고 기계학습 분야에서의 오랜 기간동안 이루어진 연구 결과의 산물이다. 어느 특정한 상황에 적용하는 특정한 기술들의 선택은 구현되어야 하는 데이터 마이닝 임무의 성격과 가용한 데이터의 성격에 의존한다. 데이터 마이닝에는 여러 임무가 있으며, 그 중에서 가장 대표적인 임무가 분류라고(classification) 볼 수 있다. 분류는 인간 사고의 기본적인 요소이기 때문에 여러 응용 분야에서 많은 연구가 진행되어 왔으며, 문제 분석의 첫 단계라고 볼 수 있다. 본 논문에서는 학습문제에서 강건성(robust)을 갖는 유전자 알고리즘 기반의 분류시스템을 제안하고, 데이터 마이닝에서 중요한 분류기능에 관련된 문제인 nDmC에 응용해서 그 유효성을 검증한다.

Abstract

Data mining means a process of nontrivial extraction of hidden knowledge or potentially useful information from data in large databases. Data mining algorithm is a multi-disciplinary field of research; machine learning, statistics, and computer science all make a contribution. Different classification schemes can be used to categorize data mining methods based on the kinds of tasks to be implemented and the kinds of application classes to be utilized, and classification has been identified as an important task in the emerging field of data mining. Since classification is the basic element of human's way of thinking, it is a well-studied problem in a wide variety of applications. In this paper, we propose a classifier system based on genetic algorithm with robust property, and the proposed system is evaluated by applying it to nDmC problem related to classification task in data mining.

1. 서 론

데이터 마이닝(Data Mining:DM)은 대용량의 데이터베이스에 숨겨져 있는 데이터간의 관계를 패턴의 탐색에 의해서 의미 있는 정보로 변환하는 작업이다. 따라서 이런 작업으로 추출된 지식은 기업의 의사결정 과정을 지원하는 등 여러 분야에서 광범위하게 활용된다. 데이터 마이닝 방법은 마이닝의 목적에 따라 분류(classification), 클러스터링(clustering), 요약(summarization) 등으로 분류되어 질 수 있다.

분류는 인간 사고의 기본적인 요소이기 때문에, 여러 응용 분야에서 분류에 관한 문제를 풀기 위한 방법들에 관해 오래 전부터 많은 연구가 진행되어 왔다. 특정한 분류 작업의 핵심은 어떻게 적당한 결정 규칙(decision rule)을 정하느냐에 달려있다. 알고리즘적인 해법에 매우 적합한 문제들에 관해서는 많은 결과와 툴이 통계와 패턴 인식 분야에 의해서 제공되어 왔으며, 또한 인간과 같은 사고 능력이 요구될 때는 인공지능분야에 관련된 방법들이 해법을 제공한다. 기계학습(machine

본 연구는 2000년도 경원대학교 학술연구비의 지원을 받아 이루어졌음.
* 정회원 : 경원대학교 공과대학 전자계산학과
mmhan@mail.kyungwon..ac.kr

learning)은 분류기(classifier)를 구축하는 작업에 다양하고 폭 넓은 접근 방법, 즉 기호적인 방법(symbolic) [1], connectionist 방법 [2], 강화기반 방법(reinforceement-based) [3], 그리고 유전자기반 방법(genetic-based)등이 있다.

유전자 알고리즘(Genetic Algorithms:GAs)은 효율적이고 독립적인 탐색 방법이며, 이미 학습하는 분류 규칙에 사용되어 왔다[4][5][6]. GA는 또한 개념 학습 [7], 특징 선택 [8], 파라미터 조정 [9], 그리고 특징 구성 [10] 에 적용되어 왔다.

유전자 알고리즘은 선택적 도태나 돌연변이 같은 생물 진화의 원리를 모델링한 확률적인 탐색 방법으로 완전검색이 불가능할 정도로 큰 탐색 공간을 갖는 최적화나 학습문제에 적용할 수 있다. 진화와 자연선택은 수백만 년을 거치면서 환경에 매우 적합한 종과 개체들로 분화되고, 적응할 수 있는 결과를 나타내었다. 이 과정들 즉, 진화와 자연선택은 한 세대에서의 가장 적합한 개체들을 유전자 물질의 번식을 통해 다음세대를 거치면서 개체들의 적합도를 최적화 하는데 도움을 준다. GA는 해결책이 개체로써 표현될 수 있다는 문제들에 대한 아이디어와, 문제는 개체들의 적합도를 최대화(최소화)하는 것이다라고 간주된다. 일반적인 유전자 알고리즘의 구조는, 순환 t동안에 해가 될 가능성이 있는 것(염색체 벡터), $P(t) = \{ x_1^t, ..., x_n^t \}$의 개체집단을 유지한다. 각 해 x_i^t는 평가되어 적합도의 척도를 준다. 그러면 더 적합한 개체들을 선택함으로써 새로운 개체집단(과정 t+1)이 구성된다. 이 새로운 개체집단의 어떤 개체들은 교배(Crossover)와 돌연변이(Mutation)에 의해 변경과정을 겪어 새로운 해를 구성한다. 교배는 부모 염색체(parents chromosome)의 일부분을 서로 바꿈으로써 부모의 특징을 결합하여 두 개의 유사한 자손들을(offsprings) 구성한다.

본 논문에서는 DM 분야에서 GA를 기반으로 하는 분류시스템을 설명하고, 분류에 관련된 문제인 nDmC에 적용해서 그 유효성을 검증하려한다. 논문의 구성은 2장에서 DM과 GA를 간략히 설명한 후 그 관련성을 설명하고, 3장에서는 실험과 그 결과를 분석한다. 마지막으로 4장에서 결론과 앞으로의 연구 방향을 논한다.

2. 데이터 마이닝과 유전자 알고리즘

2.1 데이터 마이닝

DM 또는 지식 발견(Knowledge Discovery : KD)는 일반적으로 데이터베이스에 있는 방대한 양의 정보로부터 숨어있는 지식을 자동적으로 추출하는 과정이다[11][12]. 즉 DM은 방대한 데이터베이스로부터 숨어있는 예측 정보의 추출이라고 설명할 수 있다. 이러한 정의에서처럼 방대한 원시 자료에서 유용한 정보 또는 지식을 추출하기 위해서는 일반적으로 여러 단계를 거치게 된다.

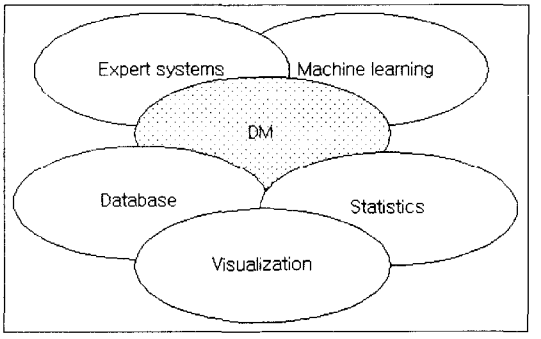
먼저 선택 단계에서 어떠한 범주에 의해서 자료를 선택하거나 분할하고, 사전처리 단계에서 어떤 정보가 이동되는 자료를 정화한다. 변형 단계에서 자료가 단순하게 이동된 것이 아니라 변형이 되어서 사용되거나 탐색할 수 있게 된다. 자료로부터 어떠한 사실들의 집합인 패턴을 추출하는데 관심이 있는 DM단계를 거쳐 최종적으로 해석과 평가 단계에서 식별된 패턴들을 인간의 의사결정 지원에 사용될 수 있는 지식 - 업무의 예측과 분류 그리고 데이터베이스의 내용 요약 또는 관찰된 현상의 설명 - 으로 해석되어진다.

지난 20여 년 간 전자적인 형태로 저장되어있는 정보와 자료의 양이 엄청나게 증가를 해왔으며, 자료의 축적은 폭발적인 비율로 발생을 해왔다. 전세계적으로 정보의 양이 20개월마다 배로 증가하고 있고, 데이터베이스의 크기와 수도 매우 빠르게 증가하고 있다. 과학적 자료 수집의 발전이 자료의 증가에 영향을 미쳤다. DM 기법을 비즈니스 의사결정의 최전방으로 이끌어 낸 여러 요인들은, 방대한 데이터베이스의 이용되지 않은 가치, 단일 고객의 관점을 지향하는 데이터베이스 레코드의 강화, 데이터베이스 통합으로부터 정보

또는 데이터 웨어하우스 개념, 하드웨어 시스템의 성능 대비 비용의 급격한 감소, 그리고 증가하고 있는 통합되는 시장에 따른 경쟁의 심화를 들 수 있다.

DM이란 이름은 방대한 데이터베이스에서 가치 있는 자료를 검색하는 것과 가치 있는 광맥을 찾기 위해서 산을 채굴하는 것 사이의 유사점에서 유래되었다. 충분한 크기와 품질의 주어진 데이터베이스로부터 DM 기술은 다음의 능력을 제공함으로써 새로운 비즈니스 경쟁력을 창출할 수 있도록 한다. 즉, 추세와 행위의 자동적 예측, 과 미리 알려져 있지 않은 형태들을 자동적인 발견이다. DM 기법들은 연구의 오랜 과정과 제품개발의 결과로써 나타나게 되었다. 이러한 발전은 비즈니스 자료들이 처음으로 컴퓨터에 저장될 때 시작이 되었고, 자료의 접근방법들이 계속적으로 발전을 했으며 더 최근에는 사용자들이 실시간으로 그들의 자료를 완전하게 탐색하도록 도와주는 기술들을 만들어 내었다. DM은 과거자료 접근과 탐색에서 미래자료와 선행 정보 전달로의 발전과정을 거치고 있다. 현재 충분하게 성숙한 다음의 기술들에 의해서 DM은 비즈니스 영역에 있어 좋은 응용을 제공할 수 있다. 즉, 방대한 자료의 수집, 강력한 멀티프로세서 컴퓨터, 그리고 DM 알고리즘이다.

DM은 귀납 학습(inductive learning), 기계 학습과 통계학 등과 같은 분야에서 도출되었다.



(그림 1) 데이터 마이닝과 다른 분야와의 관계

귀납 학습은 자료로부터 정보를 추론하는 것이고, 귀납 학습은 패턴을 발견하는 관점으로서 환경 즉 데이터베이스가 분석하는 곳에서 프로세스를 구축하는 모델이다. 보여지지 않은 객체들의 클래스를 예측하는 것을 가능케 함으로써 비슷한 객체들은 클래스들로 그룹화 되고 규칙들은 형식화된다. 그렇기 때문에 귀납은 패턴의 추론이다. 귀납 학습 방법에 의해서 만들어진 모델의 질은 모델이 미래 상황의 산출물을 예측하는데 사용되어질 수 있느냐에 달려있다. 문제점은 대부분의 환경이 다른 상태를 가지고 있다는 곳이다.

통계학은 확고한 이론적인 기초를 가지고 있지만 통계의 결과들은 거부할 수 없으며, 어디에서 그리고 어떻게 자료들을 분석하는지에 관한 사용자 지침으로써 해석하기가 어렵다. 그러나 DM은 전문가들의 자료에 관한 지식과 컴퓨터의 발전된 분석기법들을 함께 사용할 수 있도록 한다.

기계 학습은 학습 프로세스의 자동화이고 학습은 환경 적인 상태와 변천의 관찰에 기초한 규칙을 구축하는 것이다. 기계 학습은 예제를 통한 학습뿐 아니라 학습 강화, 선생님과의 학습 등을 포함하고 있는 광범위한 영역이다. 학습 알고리즘은 자료의 집합과 정보를 입력자료로 하여 학습의 결과를 표현하는 개념을 산출물로써 돌려준다. 기계 학습은 이전의 예제와 그것들의 결과를 검토하고 어떻게 그것들을 재생산하고 새로운 경우에 관한 일반화를 만들어 내는지를 배운다.

DM 방법은 그들이 수행하는 기능 또는 사용되어지는 응용의 클래스에 의해서 분류되어지며 대표적인 기능에는 분류, 평가, 예측, 친화적인 그룹화, c군집, 그리고 묘사 등이 있다. 또한 DM은 다양한 기법을 사용하고 있다. 일반적으로 더 많은 기법을 사용할수록, 더 정확한 결과를 도출할 수 있다. 이것은 하나의 기법이 의미 있는 무엇인가를 발견하지 못하면 다른 기법이 발견할 것이라는 것이며, 대표적인 기법에는 시장 바구니 분석(market basket analysis), 메모리 기반 추론(memory based reasoning), 군집 탐지(cluster detection), 링크

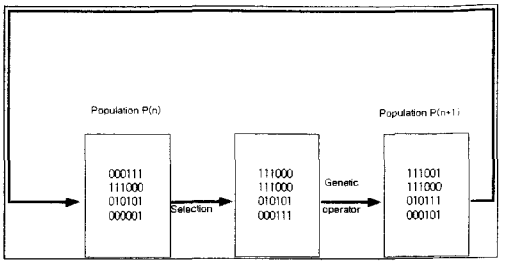
분석(link analysis), 결정 트리(decision trees), 인공 뉴럴 네트워크(artificial neural networks), 유전자 알고리즘, 그리고 온라인 분석 처리(OLAP: on-line analytic processing)등이 있다.

2.2 유전자 알고리즘

GA는 유전적 계승과 다윈적 생존 경쟁이라는 자연 현상을 모델링한 확률적인 탐색방법으로 유전검색이 불가능할 정도로 큰 후보해 공간을 갖는 최적화문제에 적용할 수 있다. 즉, 해가 될 가능성이 있는 개체집단을 유지함으로써 여러방향의 탐색을 실행하고 이들 방향간의 정보 형성과 교환을 행한다. 개체집단은 진화과정을 모방하는데, 각 세대에서 비교적 우량한 해들이 재생산되고, 반면에 비교적 불량한 해들은 소멸된다. 또한 다른 해들간의 차이를 구별하기 위해 환경의 역할을 수행하는 목적함수를 사용한다. 이러한 유전자 알고리즘은 특정한 문제에 대해 다섯 가지의 요소를 가져야만 한다. 유전자적 표현방법, 초기 개체집단을 만들어 내는 방법, 목적함수, 유전 연산자, 그리고 여러 가지 매개변수의 값이다.

어떤 개체집단을 초기화하기 위해서는 단순히 개체집단의 염색체를 비트 단위로 임의로 설정할 수 있다. 혹은 가능한 최적값 들의 분포에 관한 지식을 가지고 있다면 초기의 해집합을 배열하는데 그 정보를 이용할 수 있다.

알고리즘의 나머지 부분은 각 세대에서 각각의 염색체를 평가하고, 적합도 값에 기초한 확률분포에 의하여 새로운 개체집단을 선택하며, 돌연변이와 교배연산자에 의하여 새로운 개체집단의 염색체들을 변화시킨다. 여러 세대 후에 더 이상의 개선이 없으면, 그 세대의 가장 좋은 염색체가 최적해를 나타낸다. 선택과정에서는 적합도에 비례해서 가장 좋은 염색체는 더 많이 복제되고, 보통 염색체는 비슷하게 남아 있으며, 최악의 염색체는 소멸된다. 교배연산자는 교배연산확률을 토대로 두개의 염색체에 적용되어서 새로운 두 개의 자손을 생산하며, 마지막으로 돌연변이 연산자가 돌

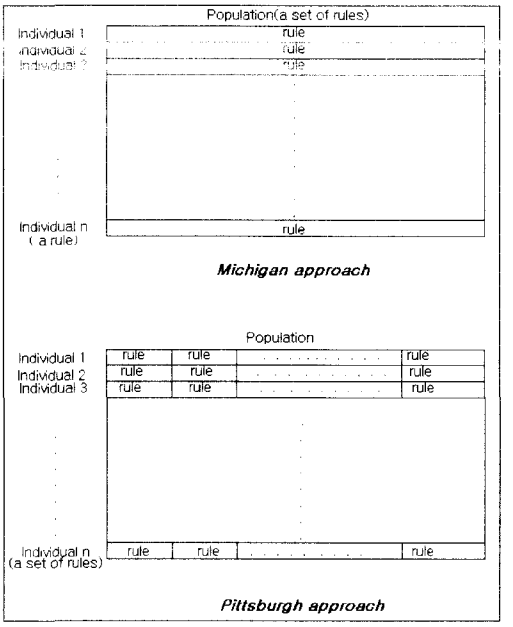


(그림 2) 유전자 알고리즘의 예

연변이 확률에 의해 비트별로 적용된다. 이러한 선택, 교배, 그리고 돌연변이를 한 후에 새로운 개체집단은 다음 평가를 받는다. 유전자 알고리즘의 기본 동작은 그림 2과 같다.

2.3 유전자 알고리즘의 데이터 마이닝에의 접근

대부분의 데이터 마이닝 시스템은 전통적인 기계학습알고리즘의 변형을 사용해 왔다. 기계학습에서는 복잡한 시스템을 대상으로 하여 그 대상 시스템을 학습시킬 뿐만 아니라 시스템에 대한 적절한 출력을 만들어 내는 두 가지의 목적을 가진다. 기계학습에서 유전자 알고리즘의 기법을 이용한 것을 GA기계학습 또는 GBML(genetic based machine learning)이라고 한다. 기계학습이 최적화 문제와 근본적으로 다른 점은 규칙의 조를 구하지 않으면 안 되는 점이다. 최적화 문제에서는 최적화에 가까운 우수한 해를 구하는 것이 목적이기 때문에 최후의 한 종류만이 개체에 수렴하면 되지만 기계학습에서는 가장 좋은 규칙 하나만 구하는 것이 아니라 서로 협조하는 규칙의 집합을 구하는 것이 필요하다. GBML에서는 일반적으로 두 가지 접근 방법이 있다. 전체 규칙 집합을 하나의 개체로 표현하고, 후보 규칙 집합들의 개체 집단을 유지하고, 그리고 규칙 집합들의 새로운 세대를 생성하기 위한 선택과 유전 연산자를 사용하는 것이 자연스러운 방법으로 여겨질 것이다. 즉, 전통적인 유전자 알고리즘을 사용하며, 집단 안에서의 각 실재(entity)는 학습 문제에 대한 완전한 해를 표현하는 규칙의 집합이다. 이러한



(그림 3) Michigan 과 Pitt 접근방법

접근 방법을 Pitt 접근 방법이라 한다[13]. 또한 같은 시기에 Holland는 개체 집단의 소속원 들이 각각의 규칙들이고, 하나의 규칙 집합이 전체 개체 집단에 의해 표현되는 분류 시스템을 발전시켰다. 이러한 방법은 Michigan 접근 방법이라 불리워진다. 이러한 Michigan 접근 방법은 상당히 다른 진화 방법을 사용하는데, 집단은 개인적인 규칙들로 구성되었으며 각 규칙은 전반적인 학습 임무에 대한 부분 해를 표현한다[14]. 즉 Pitt 접근 방법은 진화 연산과 유사하지만, Michigan 접근 방법은 매우 다른 새로운 방법을 사용한다.

3. 실험 및 고찰

3.1 문제 영역

실험은 4개의 특징과 각 특징은 4개의 가능한 값으로 구성된 인공적으로 만들어진 문제에서 실행했다. 따라서 256개의 경우(instance)가 존재한다. 이 문제에서 12개의 목적 개념(target concept)을 구성했으며, 목적 개념의 복잡도를 개념들을 정확히 기술하는데 요구되는 규칙당의 규칙의 수(disjuncts)와 관련된 특징의 수(conjuncts)를 증가시킴으로서 변동했다. 선언의 수는 1부터 4까지이며, 결합의 수는 1부터 3까지이다. 각 목적 개념은 nDmC로 표시되는데, 여기서 n은 선언의 수이고 m은 결합의 수이다 [15].

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

모든 목적 개념은 다음의 형태를 가진다.

4DmC == d1 $\vee$ d2 $\vee$ d3 $\vee$ d4

3DmC == d1 $\vee$ d2 $\vee$ d3

2DmC == d1 $\vee$ 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)

또한 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)

마지막으로 nD1C 목적 개념은 다음과 같다.

d1 == (F1 = v1)

d2 == (F1 = v2)

d3 == (F1 = v3)

d4 == (F1 = v4)

목적 개념들의 각각을 위하여, 집합에 있는 모든 256개의 경우들은 목적 개념의 긍정 혹은 부

정으로서 표시한다.

3.2 탐색공간의 표현

개념들을 표현하는 자연스런 방법은 분류 규칙들을 가능한 중복하는 선언 집합(disjunctive set), 즉 선언 상태 형식(disjunctive normal form:DNF)이다. 각 규칙의 왼쪽은 특징의 값을 포함하는 하나 혹은 그 이상의 테스트들의 결합으로 구성된다. 규칙의 오른쪽은 규칙의 왼쪽에 의해서 부합되는 예제에 할당되고 있는 개념이나 분류를 표현한다. 즉, 만약에' 규칙들이 특징 공간의 요소들을 정확히 분류한다면 이러한 규칙들의 집합은 알려지지 않은 개념을 표현하는 것으로 생각할 수 있다.

이러한 규칙들의 결합적인 왼쪽에 임의의 복잡한 항목을 허용한다면, 스트링으로 표현하기 어려운 매우 강력한 표현 언어를 가져야 한다. 그러나, 결합의 요소들의 복잡도를 줄임으로서, 좀더 많은 규칙들이 개념을 표현하는데 필요하다는 단점을 갖는, 표준 GA와 스트링 표현을 사용할 수 있다. 즉 형태의 테스트 되어지는 결합의 각 요소를 줄임으로서 얻게 된다.

예를 들면, 규칙은 다음과 같은 심볼 형태를 갖을 수 있다.

만약에 (F1 = large)와 (F2 = sphere 혹은 cube) 이면 그것은 연장이다.

위의 규칙에서 왼쪽이 내부가 선언을 갖는 결합형태이기 때문에, 각 특징에 많아야 한 테스트가 있어야하는 일반성을 잃지 않는다. 이런 규약을 가지고 분류 규칙을 위한 고정된 길이의 내부 표현을 만들 수 있다. 각 고정된 길이의 규칙은 각 특징에 한번씩의 N 특징 테스트를 갖는다. 각 특징 테스트는 고정된 길이의 이진 스트링에 의해서 표현되며, 그 길이는 특징의 특성에 의존한다. 이 논문에서는 명사형 값(nominal value)만을 갖는 특징을 사용한다. 시스템은 명사형 특징의 k 값을 위하여 k 비트를 사용한다. 예를 들어, 특징 F1의 값의 집합이 {small, medium, large} 이면, 011은 medium 혹은 large인 F1을 위한 테스트를 표현한다. 또한 특징 F2가 {sphere, cube, brick, tube}의 값을 갖고, 연장과 부품의 두 클래스가 있다고 가정한다. 그러면 이 두 특징의 문제에 대한 규칙은 다음과 같이 표현된다.

F1	F2	Class
111	1000	0

이 규칙은 다음과 같다.

만약에 (F1 = small 혹은 medium 혹은 large) 와 (F2 = sphere)이면 그것은 연장이다.

특징의 모든 값이 1인 특징 테스트는 그런 항목을 누락하는 것과 같으므로 위의 규칙은 다음과 같다.

만약에 (F2 = sphere)이면 그것은 연장이다.

3.3 분류 규칙들의 집합

개념기술(concept description)은 하나나 그 이상의 분류 규칙으로 구성되었기 때문에 GAs가 규칙의 집합들을 진화시키기 위해 사용된 방법을 기술 할 필요가 있다. 앞에서도 설명했듯이 기본적으로 두 종류가 있다. Holland의 분류 시스템(classifier system)으로 대표되는 Michigan 접근 방법과 Smith의 LS-1시스템으로 대표되는 Pittsburgh 접근 방법이 있다. Michigan 접근 방법을 사용하는 시스템들은 집단안에서 공간과 우선순위를 위해 서로 경쟁하는 개인 규칙의 집단을 유지한다. 그에 반해, Pittsburgh 접근 방법을 사용하는 시스템들은 문제 영역에서 수행 능력에 따라 서로 경쟁하는 길이가 변하는 규칙 집합의 집단을 유지한다. 이 논문에서는 Pittsburgh 접근 방법을 사용해서 얻은 결과를 보고한다. 즉, 집단의 각 개체

는 고정된 길이의 규칙의 집합을 표현하는 길이가 변하는 스트링이다. 특정한 개체 안에 있는 규칙들의 수는 제한이 없든가 혹은 사용자가 상한치를 정한다.

이 표현을 좀 더 구체적으로 설명하기 위해, 다음의 두 개의 규칙을 가진 규칙 집합의 예를 고려한다.

F1	F2	Class	F1	F2	Class
100	1111	0	011	0010	0

이 규칙의 집합은 다음과 같다.

만약에 (F1 = small)이면 그것은 연장이다.
혹은
만약에 (F1 = medium 혹은 large) 와 (F2 = brick)이면 그것은 연장이다.

3.4 유전자 조작과 적합함수(Genetic operator and Fitness function)

유전자 조작(genetic operator)은 측정과 평가를 위해 새로운 개체들을 생산하기 위하여 집단안에서 개체들을 수정한다. 전통적으로 교차와 돌연변이가 가장 중요하다. 교차는 두 개의 개체를 얻어 두 개의 새로운 개체를 유전 자료의 부분을 바꿈으로서 생산한다. 돌연변이는 집단안에 있는 비트들을 작은 확률의 바탕으로 바꾼다. 즉 교차는 부모의 염색체 일부분을 교차율을 바탕으로 서로 바꿈으로써 두 개의 유사한 자손을 구성하며, 돌연변이 는 돌연변이율과 동일한 확률을 가지고 임의로 변경시킨다. 이러한 기본적인 조작들을 이용할 수 있는 개념 학습 표현은 앞 절에서 기술한 방법으로 달성할 수 있다.

실험에서 사용한 돌연변이 조작은 전통적인 것을 사용했으며 비트 단위에서 실행한다. 교차는 규칙 집합의 길이가 변하는 것을 다루기 위해서 전통적인 2점 교차에서 일반적으로 확장된 방법을 사용한다. 고정된 길이의 스트링에서 일반적인 2점 교차는, 교차점은 항상 두 부모에 맞춰야 하기 때문에, 교차점을 선택하는데 두 단계의 자유만이 있다. 그러나, 길이가 변하는 스트링에서는 첫 번째 부모에서 2점 교차점을 선택해서 두 번째 부모에서 같은 점이 존재하리라는 보장이 없기 때문에 네 단계의 자유가 있다. 따라서 교차점의 두 번째 집합은 이것을 고려해서 선택되어야만 한다.

표준 교차에서는 어디서 교차점이 일어나는지에 제한이 없다. 오직 고려할 점은 두 부모상에서 해당되는 교차점이 의미적으로 맞아야 한다는 것이다. 즉, 한 부모가 규칙의 경계 점에서 짤리면 다른 부모도 규칙의 경계 점에서 짤려야만 한다. 같은 방법으로 한 부모가 규칙의 경계점 오른쪽에서 5 비트점에서 짤리면, 다른 부모도 같은 점에서 짤려야 한다.

예를 들면, 다음의 두 규칙집합들을 고려하자.

F1	F2	Class	F1	F2	Class
100	0100	0	011	0010	0
010	0001	0	110	0011	0

왼쪽 짤리는 점이 규칙 경계 점에서 두 비트가 들어갔고, 오른쪽 짤리는 점은 규칙 경계 점에서 한 비트가 들어간 것을 주의하자. 짤리는 점안에서 비트들을 바꾸면, 다음과 같은 세 개의 규칙들의 규칙 집합과 한 규칙의 규칙 집합이 된다.

F1	F2	Class	F1	F2	Class	F1	F2	Class
100	0001	0	110	0011	0	011	0010	0
010	0100	0						

좋은 표현을 선택하고 난 후, 올바른 종류의

개체에 보상해주는 좋은 적합함수를 정의하는 것이 중요하다. 이 논문에서는 오직 분류 성능만을 포함하는 적합함수를 선택했다. 각 개체 규칙 집합의 적합도는 훈련 예제(training example)의 현재 집합에서 규칙 집합을 검사함으로서 계산되어진다.

fitness(individual i) = (percent correct)^2

3.5 유전자 알고리즘 기반의 개념 학습자

이 논문에서 사용한 시스템, 유전자 알고리즘 기반의 개념 학습자, 을 설명한다. 이 시스템의 핵심은 긍정과 부정 예제의 주어진 집합에서 잘 수행하는 규칙을 위해 규칙 집합의 공간을 탐색하는 GA이다.

```
procedure GA;
    begin
      t = 0;
      initialize population P(t);
      fitness P(t);
      until(done)
          t = t + 1;
          select P(t) from P(t-1);
          crossover P(t);
          mutate P(t);
          fitness P(t);
      end.
```

(그림 4) 시스템과 GA

P(t)는 규칙 집합들의 집단을 표현한다. 집단을 임의로 초기화 한 후, 각 규칙 집합은 앞에서 설명한 적합 함수에 의해 평가 된다. 규칙 집합들은 그들의 적합도에 비례해서 확률적으로 선택되어진다. 교차와 돌연변이가 확률적으로 살아 남은 규칙 집합에 적용되고, 새로운 집단이 생성된다. 이러한 사이클이 시간/공간의 주어진 제약안에서 지속적이고 완벽한 규칙집합이 발견될 때 까지 계속된다.

이러한 시스템들은 모든 경우가 한번에 시스템에 나타나는 배치 모드와 어떤 시간에 하나나 몇몇의 경우가 시스템에 나타나는 점진적 모드의 두 종류 형태로 나눌 수 있다.

3.6 결 과

위에서 언급한 시스템의 성능을 제안된 문제영역에서 평가했다. 실험에 사용된 파라미터 값으로 2점 교차율은 0.6으로 정했으며, 돌연변이율은 0.001을 사용했다. 집단크기를 정하는 것은 다소 어려운 일이다. 탐색공간이 크고 복잡한 경우에는 집단크기가 상대적으로 큰 것이 요구되지만 실행시간이 더 요구된다. 이 실험에서는 1000개의 집단크기를 정했다.

3.7 고 찰

특정 경우들로부터 일반적인 사실을 이끌어 내는 방법을 귀납이라고 한다. 즉, 많은 예제로부터 그것들의 일반적 경향을 알아내는 작업인데, 이러한 경향을 도출해 낼 수 있다면 새로운 예제에서 어떤 결과를 예측할 수 있게 된다. 이러한 분야의 시스템들은 여러 방식이 있는데, 심벌한 방법(C4.5 등), 뉴럴 네트워크을 사용한 방법, 확률을 사용한 시스템, 그리고 GA를 사용한 방법등이 있다. 이 논문에서는 Pittsburgh 접근 방법을 사용하였다.

이 시스템에서는 규칙 표현이 1 과 0 으로 구성되어 있다. 여기서 1이란 어떤 속성을 지니고 있는 규칙을 나타내며, 0이란 규칙에 있어서 어떤 속성을 고려하지 않겠다는 의미이다. 예를 들면, 어떤 규칙의 속성 1이 "축구를 좋아한다."이고, 속성 2가 "야구를 좋아한다." 일 때 10 이란 규칙은 "축구를 좋아한다." 라는 것을 나타내게 된다.

위에서 본 바와 같이 nDmC 문제에서 규칙을 올바르게 생성했으며 수행 능력도 우수하다고 볼 수 있다.

4. 결 론

데이터 마이닝은 데이터에 내재되어 있는 유용한 패턴이나 변수들간의 관계를 정교한 분석 모형을 사용하여 찾아내는 방법이다. 데이터웨어하우스를 기반으로 한 필수적인 분석요소인 데이터 마이닝 분야는 인공지능의 분석기법이 인식을 비롯한 인지 과학분야에서의 성공적 기반을 바탕으로, 대규모의 정보홍수와 치열한 경쟁환경에 직면한 기업환경에서의 전략적 의사 결정지원을 위한 분야이다. 즉, 대규모의 전략적 데이터베이스로부터 숨겨진 미래에 대한 예측정보를 추출하는 일련의 과정이며 데이터 웨어하우스상에 내재되어 있는 기업에 있어서의 가장 소중한 입력정보로부터 전략적 판단을 위한 분석 정보를 추출하는 강력하고도 새로운 적용분야이다.

(표 1) 제안된 시스템에 의한 수행능력

problem	gen	generated rule	performace
1D1C	1	1000 1111 1111 1111	100%
1D2C	7	1000 1000 1000 1111	100%
1D3C	45	1000 1000 1000 1110	100%
2D1C	1	1111 1000 1111 1111	
	7	0100 0100 1111 1111	100%
2D2C	6	1000 1000 1111 1111	
	7	0100 0100 1111 1111	100%
2D3C	90	0100 0100 0100 1111	
	14	1000 1000 1000 1111	100%
3D1C	1	1111 1111 1000 1111	
	1	1000 1111 1111 1111	
	1	1111 1000 1111 1111	100%
3D2C	19	1000 1000 1111 1111	
	18	0010 0010 1111 1111	
	7	0100 0100 1111 1111	100%
3D3C	153	1000 1000 1000 1111	
	50	0100 0100 0100 1111	
	373	0010 0010 0010 1111	100%
4D1C	1	1111 1111 1000 1111	
	1	1111 1111 1111 1000	
	1	1111 1000 1111 1111	
	1	1000 1111 1111 1111	
4D2C	10	0001 0001 1111 1111	100%
	21	1000 1000 1111 1111	
	61	0100 0100 1111 1111	
	35	0010 0010 1111 1111	
4D3C	549	1000 1000 1000 1111	100%
	24	0010 0010 0010 1111	
	124	0100 0100 0100 1111	
	20	0001 0001 0001 1111	

이 논문에서는 유전자 알고리즘을 사용하여 강건한(robust) 개념 학습 시스템을 제안했으며, 이러한 방법으로 데이터 마이닝의 핵심적인 기능인 분류를 Pittsburgh 기법을 활용하여 실행했다. 제안된 시스템은 병렬 처리가

가능해서 실행 속도를 향상시킬 수 있으며, 다른 특정한 시스템과의 연계가 용이해서 수행성능을 높일 수 있다.

또한 폭 넓은 범위에 응용해서 좋은 결과가 기대된다.

장차 연구 할 방향으로서 수행능력을 향상시킬 새로운 유전자 조작이나 방법에 대한 연구가 필요하며, 또한 다른 문제 영역에서의 적용을 검토하고 있다.

참고문헌

[1] R. Michalski, "A Theory and Methodology of Inductive Learning," R. Michalski, J. Carbonell, and T. Mitchell, eds., Machine Learning: An AI Approach, vol. 1. Morgan Kaufmann, Los Altos, Calif., pp.83-134, 1983.

[2] D. E. Rumelhart, and J. L. McClelland, "Parallel Distributed Processing:Exploration in the Microstructure of Cognition," PDP Research Group eds., Cambridge, Mass.: MIT Press, 1986.

[3] R. S. Sutton, "Learning to Predict by the Methods of Temporal Differences," Machine Learning, vol. 3, pp.9-44, 1988.

[4] D.E. Goldberg, Genetic Algorithms. Addison-Wesley, 1989.

[5] J.H. Holland, "Adaptation in Natural and Art-

ificial Systems," Ph.D. Thesis, Univ. of Michigan, Ann Arbor, Mich. 1975.

[6] S.Wilson, "Classifier Systems and the Animat Problem," Machine Learning, vol. 2, 199-228, 1987.

[7] J. Bala, K.A. De Jong, and P. Pachowicz, "Learning Noise Tolerant Classification Procedures by Integrating Inductive Learning and Genetic Algorithms", Proc. First Int'l Workshop on Multistrategy Learning, Harpers Ferry, W. Va., pp.316-323, 1991.

[8] F.Z. Brill, D.E. Brown, and W.N. Martin, "Fast Genetic Selection of Features for Neural Network Classifiers,"IEEE Trans. Neural Network, vol. 3, no. 2, pp.324-328, 1992.

[9] M.Botta, A.Giordana, and L.Saitta, "Learning Fuzzy Concept Definitions," Proc. Second IEEE Int'l Conf. Fuzzy Systems, San Francisco, Calif., pp.18-22, 1993.

[10] A.Giordana, G.Lo Bello, and L.Saitta, "Abstraction in Propositional Calculus," Proc. Workshop Knowledge Compilation and Speed Up Learning, Amherst, Mass., pp.56-64, 1994.

[11] Michael J. A. Berry and Gordon Linoff, Data Mining Techniques For Marketing, Sales, and Customer Support, Wiley Computer Publishing, 1997.

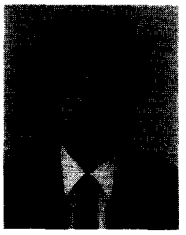
[12] Pieter Adriaans and Dolf Zantinge, Data Mining, Addison-Wesley, 1996.

[13] Stephen F.Smith, A Learning System Based on Genetic Adaptive Algorithms. Ph.D. thesis. University of Pittsburgh, 1980.

[14] John H.Holland, Escaping brittleness: the possibilities of general purpose learning algorithms applied to parallel rule-based systems, Machine Learning, an artificial intelligence approach, 2. 1986.

[15] Kenneth A.De Jong, William M.Spears, and Diana F.Gordon, Using Genetic Algorithms for Concept Learning, Machine Learning, 13, pp. 161-188, 1993.

◐ 저 자 소 개 ◑



한 명 묵

1980년 연세대학교 요업공학과 졸업(학사)
1987년 뉴욕공과대학교 전자계산학과 졸업(석사)
1997년 오사카시립대학교 정보공학과 졸업(박사)
현 재 경원대학교 전자계산학과 교수
관심분야 : 알고리즘, Data Mining, Pattern Recognition