韓國情報敎育學會論文誌
第3卷 第1號 pp.75~85

교육에서의 효율적인 정보 활용을 위한 데이터 마이닝 기법

이철환 · 한선관
인천교육대학교 컴퓨터교육과
chlee56@pine.inchon-e.ac.kr • fish@compedu.inchon-e.ac.kr

요약

본 연구는 초·중등교육에서 사용되고 있는 데이터 베이스 시스템에 데이터 마이닝 기법을 적용하여 보다 효율적인 교육자료로 활용하기 위한 방안 제시에 그 목적이 있다. 데이터 마이닝에 대한 전반적인 내용과 기계학습과 관련된 내용을 고찰하였다. 교육에서 많이 사용되는 데이터베이스 시스템으로 종합생활기록과 건강 기록, 성적 자료가 있으며, 이러한 자료에서 나타난 특별한 형식과 집합을 데이터 마이닝 기법과 기계학습을 이용하여 유용한 정보를 추출하는 방법에 대해 제시하였다. 그리고 이러한 데이터 마이닝 기술을 사용함에 있어 교육 현장에서 문제가 되는 점과 이를 해결하기 위한 방안을 제안하였다.

Data Mining Technology for Efficient Information Application in Education

Chul-Hwan Lee* · Sun-Gwan Han**
Inchon National University of Education, Dept. of Computer Education

Abstract

The purpose of the paper is to apply a Data Mining method to Data Base System for more efficient educational data used in elementary and secondary education. First, this study investigated the whole contents of Data Mining and technique relation to Machine Learning. Mainly Data Base Systems in education are general life checking, record of health, and score reports. We suggested Data Mining method and Machine Learning when we search for information of usefulness in a particular representational form or a set of such representations in data. Also, we propose the problem and the solution when using data mining techniques in education.

1.서론

21세기 정보화 시대를 맞이하여 교육 현장에서는 다양한 방법으로 미래정보화 시대에 적응하려고 많은 노력과 시도가 있었다. 특히 컴퓨터 교육에 대한 관심이 고조되면서 컴퓨터 교육의 강화와 컴퓨터 기자재의 적극적인 지원, 컴퓨터 업체와의 연계를 통한 사설 컴퓨터 교육, 교사의 컴퓨터 자율 연수 강화, 교단 선진화 기자재의 보급, 학습 업무에의 적극 활용, 소프트웨어 공모전을 통한 교사의 컴퓨터에 대한 참여와 관심의 고조 등이 바로 그 내용이다.

그러나 이러한 다양한 시도에도 불구하고 문서작성이나 컴퓨터 보조 기능학습을 제외하고, 컴퓨터 공학적인 측면에서 실제적인 업무와 교육에의 적용이 매우 미흡하게 사용되는 것이 지금 교육의 현실이다. 컴퓨터의 가장 기본적이며 핵심적인 기능인 데이터의 저장과 검색을 통한 정보의 활용이 매우 드물고 국가 차원에서 교육의 기본정보를 데이터 베이스로 구축한 부분 또한 거의 없는 실정이다.

교사의 개인 인사 기록과 학생의 성적과 종합 생활 기록부, 건강 기록부는 교사와 학생에 관한 종합적인 지식을 나타낸 정보의 보고로서 교육에 다양하게 활용될 수 있으며, 대규모의 데이터 베이스로 완벽하게 구축되어야 할 절실한 부분이다. 근래에 들어 학생의 성적, 종합 생활 기록부와 건강 기록부가 전산화되고 있지만 전국적으로 체계화되어 있지 않으며 이를 교육에 적용하는 예는 극히 드물다. 교사가 전산화된 학생의 정보를 직접 찾거나 검색하여 이용하는 경우가 대부분이기 때문에 전산화 작업은 결국 교사의 업무를 더 과중하고 힘들게 하는 부분이 되었다.

이러한 대규모의 학생 전산 기록 자료는 데이터 베이스로 처리하여 교육에 적극적으로 활용하기에 매우 훌륭하며 귀한 자료이다. 그렇기 때문에 학생들의 다양한 행동과 학습 사례에 대한 상호 관련성을 발견할 수 있고, 유용한 패턴을 추출하여 다양한 부분의 교육에 적용할 수 있다.[9]

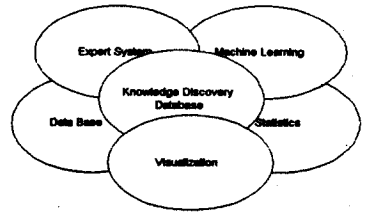
컴퓨터를 통해 많은 데이터에서 '중요한 정보를 추출하는 방법을 데이터 마이닝(Data mining) 기법이라고 하는데 다량의 데이터를 토대로 하여 추론을 하고 기계학습을 통하여 새로운 정보와 패턴을 밝혀내는 컴퓨터의 새로운 응용 분야이다.[1] 외국의 경우 네트워킹 환경의 Distributed Data Base 자료에서 Global pattern을 탐색하기 위한 방안으로 에이전트 기술을 접목한 데이터 마이닝 기술이 많이 연구개발되고 있는 실정이다.

따라서 본 연구에서는 이러한 데이터 마이닝의 다양한 기능을 고찰하고, 교육에의 적용 가능성 여부와 적용 부분을 살펴본 뒤 데이터 마이닝 적용 모델을 제시하며, 이에 따른 문제점과 해결 방안을 제시한다.

2. 이론적 고찰

2.1. 데이터 마이닝의 정의

데이터 마이닝의 정의를 살펴보면, '대규모의 DB 내에 존재하지만 숨겨져 있는 상호 관련성과 글로벌 패턴(Global Pattern)에 대한 탐색', '인공지능 기술, 패턴 인식 기술, 통계기법, 수학적 알고리즘, 인지과학을 이용하여 의미 있는 새로운 상관 관계, 패턴, 추세 등을 학습을 하여 새로운 지식(자료)을 발견하는 과정'이라고 볼 수 있다. 즉 대용량의 Data Base로부터 이제까지 몰랐던 정보를 추출하는 과정이라고 정의 내리고 있다.[11] 데이터 마이닝은 지식의 발굴(Knowledge Discovery of Data base, KDD)이라는 용어로 사용되기도 하는데, 구조화된 데이터를 통하여 새로운 지식을 탐색하고 새로운 지식을 생성하는 과정이라고 볼 수 있다. 이에 대한 효과는 정보검색이나 보고서가 밝혀 낼 수 없었던 정보를 밝혀내고, 추론을 통해 의미 있는 자료를 추출하며, 학습 기능을 통하여 새로운 지식을 구축할 수 있다.[13]



[그림 1] KDD와 데이터 마이닝

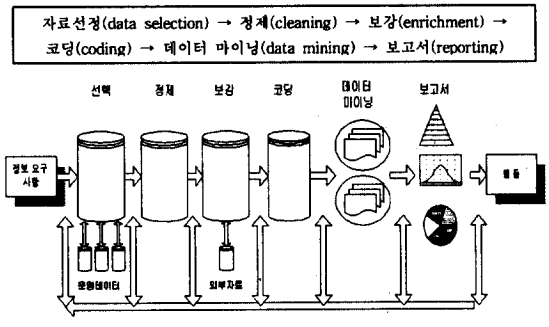
지식 발굴과 관계된 데이터 마이닝의 관계 영역을

살펴보면 [그림1]과 같이 데이터 베이스(DataBase System)를 기반으로 하여 인공지능의 전문가 시스템(Expert System of Artificial Intelligence), 기계학습(Machine Learning), 통계학(Statistics), 가시화(Visualization)기술의 분야가 응용되고 있다.[10]

데이터 마이닝은 6가지 모델로 나뉘는데 분류(Classification), 군집화(Clustering), 회귀분석(Regression), 시계열 분석(Time series analysis), 연관 규칙(Association analysis), 순차 패턴 탐색(Sequence analysis)의 기법들이 다양하게 사용되고 있다.[11]

2.2. 데이터 마이닝의 지식 탐사 과정

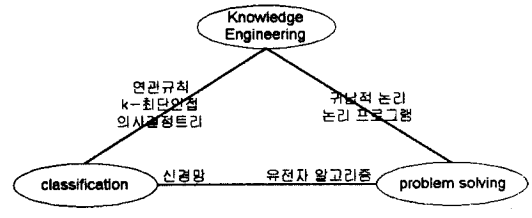
지식의 탐사과정은 일반적으로 [그림 2]와 같이 6단계를 거친다.



[그림 2] KDD 절차

기본적인 데이터 베이스 시스템에서 자료를 선정(data selection)하고, 정제(cleaning)과정에서 사전의 오류를 발견하고 중복을 제거하며, 도메인의 일관성이 결여된 경우 NULL값으로 대치한다. 보강과정(enrichment)에서 기존의 레코드 데이터와 조인하는 SQL 연산을 통해 충분한 정보만을 가진 레코드만 선정한다. 코딩(coding) 작업에서는 패턴인식 알고리즘의 입력으로 사용하기에는 너무 자세한 정보를 도메인 값의 축소를 통해 시계열 패턴을 발견한다. 이러한 과정이 모두 끝나게 되면 본격적으로 다양한 데이터 마이닝(Data Mining) 기법을 적용하여 유용한 정보를 추출해 내며 분석하고 학습하며, 결과를 보고서(Reporting)로 출력한다.[12]

기본적으로 많이 사용되는 데이터 마이닝의 기법을 살펴보면 SQL을 이용하여 표층 Data(지식) 탐색하는 Query tool, SQL의 결과값을 가지고 통계적으로 분석하는 Statistical technique, 시각화 도구를 사용하는 Visualization, 학습을 하지 못하지만 다차원 분석을 하는 OLAP (OnLine Analytical Processing), 최단 인접한 이웃의 사례를 기반으로 추론하는 CBL(Case based learning, k-nearest neighbor), 의사결정과정을 트리 구조를 통해 직관적으로 제공하는 Decision tree, 생성 규칙(IF~THEN)을 이용하는 Association tree(Association rule), 인간의 두뇌를 모델로 개발한 Neural network, 코딩 문제에 대한 기법으로 DNA를 이용한 Genetic algorithm 등이 많이 사용된다.[13] 이러한 데이터 마이닝의 기법은 그 나름대로의 장단점이 있으며 다음 [그림 3]을 보면 크게 3가지 작업으로 분류하여 알맞은 곳에 적용해야 한다.



[그림 3] 기계학습과 작업 유형

<표 1> 알고리즘 효율성

구분	처리 내용	k-최단 인접	의사결정트리	연관 규칙	신경망	유전자 알고리즘
입력 품질	많은 레코드 수 처리 능력	△	○	○	△	△
	많은 항목 수 처리 능력	△	○	△		
	수치 항목 처리 능력	○	○		○	
	스트링 처리 능력				△	△
출력 품질	규칙학습 능력		○	○		○
	점진적 학습 능력			○	△	△
	통계적 중요성 평가 능력	○	○	○		
학습 성능	디스크 부하	○	△	△	△	△
	CPU부하	○	△	△		
응용 성능	디스크 부하		○	○	○	○
	CPU부하	△	○	○	△	△

<표 1>은 각 알고리즘에 따른 효율성을 나타낸 것이다

2.3. 교육에서의 데이터 마이닝

데이터 마이닝은 기업의 마케팅과 전자 상거래 부문에서 활발하게 연구, 진행되어 오고 있으며 교육에 적용된 예는 극히 드물다. 대부분 데이터 수집과 분석, 대량의 자료, 인간 전문가를 대신하여 신속하고 정확한 자료를 추출하는 분야, 비용측면에서 인간이 처리할 수 있는 것 보다 효율적으로 처리할 수 있는 부분에서 많이 활용되고 있다.

일반적인 목적에 데이터 마이닝을 적용하기 위한 전제 조건을 살펴보면 다음과 같다.[10]

- 대용량의 데이터 베이스시스템(Larger Database System):Multi-gigabyte 또는 terabyte(10^{12})이상
- 다단계의 복잡도(High Dimensionality)-대용량의 필드(속성, 변수)
- 데이터의 엄격함(Overfitting)-최적의 탐색기법
- 통계적 검사의 평가(Assess)-기법의 선택
- 지식과 자료의 변경가능(Changing)
- 잡음 제거(Noisy data)-오류 및 중복자료 제거
- 필드간의 복잡한 관계(Complex relationships)
- 패턴의 이해가능성(Understandability of Pattern)
- 사용자의 흥미와 최우선 지식(User Interesting)
- 다른 시스템과의 통합(Integration)

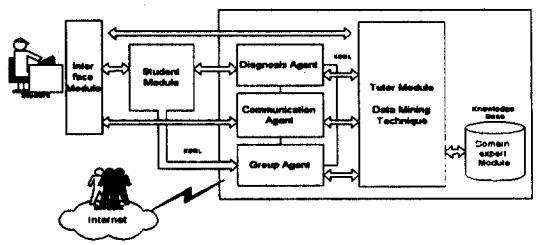
이러한 내용을 전제로 하여 교육에 접목시킬 수 있는 부분의 조건을 살펴보면 우선, 대규모의 Data Base로 구축이 가능한 자료이어야 하며, 각 Data Base가 다양한 변수와 속성으로 이루어진 필드값을 가져야 하며, 다양한 탐색 기법과 학습 Algorithm을 사용하여 지식을 추출할 수 있어야 한다. 그리고 지속적인 자료의 추가와 저장이 가능한 자료이어야 하며, 이 자료를 토대로 수정, 변경할 수 있어야 한다. 또한 결과에 대한 패턴의 양상들이 교육에 적절한 자료를 제시하고 결과가 교육에 적극 반영될 수 있는 내용이어야 한다. 또한 한 학생의 자료가 지속적으로 연결되어 저장되어야 하며, 그 자료에 대한 추론의 패턴과 인간(교사)이 인지하지 못하는 숨은 패턴을 찾을 수 있어야 한다. 비용측면에서 기존의 자료보다 그 결과가 효율적이어야 한다. 마지막으로 다른 시스템과 호환이 되어 값들이 전달, 통합될 수 있어야 하고 또한 다른 시스템과 유기적으로 사용될 수 있어야 한다.

이러한 전제 조건의 항목을 살펴볼 때 데이터 마이닝을 적용할 수 있는 부분은 학생들의 종합 생활 기록과, 건강 기록부, 성적 및 학습 진단과 교사의 인사 기록 등에 적용될 수 있다.

학생의 종합 생활 기록은 내용면에 있어서 학생의 기초자료인 인적사항, 학적사항, 출결상황, 신체 발달 자료, 심리 검사 자료, 수상경력, 자격증, 진로에 관한 자료와 학습 및 성적 자료인 교과 학습 발달 상황, 재량시간, 특별활동, 봉사활동, 행동발달 상황, 종합 의견 등 매우 훌륭하고 다양한 자료가 포함된 종합적인 학생의 정보 창고이다. 각각의 항목은 서로 유기적인 관계를 갖고 있으며 이를 학습에 적용하기에 매우 훌륭한 자료이다. 또한 전국적으로 동일한 형태의 시스템을 구축하고 있으며 전산화 작업으로 대용량의 DB가 일부분 구축되고 있는 실정이다.

건강 기록부를 살펴보면 기초 신상 자료와 전염병 예방접종 기록, 병력 기록, 구강 병리검사, 체격 및 체질 검사, 체력검사 등으로 구성되어 있으며 자료의 내용이 초등학교 1학년부터 고등학교 3학년까지 12년간의 개인의 건강과 성장에 관한 매우 중요한 자료이다. 다양한 패턴의 결과를 추출할 수 있어 학생의 성장과 질병 예방 및 진단에 다양하게 이용될 수 있다. 역시 전산화 작업으로 대용량의 DB로 구축될 예정이다.

인터넷을 이용한 가상학교의 급성장으로 원격 교육이 활성화되고 있다. 면대면(face to face) 학습에서 가상 환경으로 전환하면서 학생의 지식 수준 및 태



[그림 4] Agent 기반 데이터 마이닝 기법

도를 컴퓨터가 파악할 수 없기 때문에 대용량의 지식베이스(Knowledge base)를 서버에 구축하고, 학습

자의 학습 내용과 결과를 축적하여 그것을 기반으로 학습자의 수준과 학습에 대한 태도를 진단할 수 있도록 지능적인 에이전트 기반 데이터 마이닝 모델을 구현할 수 있다. 클라이언트와 서버간의 상호 통신은 KQML을 이용한 에이전트 시스템으로 [그림 4]의 모델과 같이 구현할 수 있다.

교사의 인사자료는 인사 발령과 근무태도, 연수자료, 건강, 승진, 학습에 관한 기초 자료 등에 적절히 이용될 수 있으나 전체적인 내용에 대해 아직 전산화가 이루어지지 않고 있지만 부분적인 내용이 각각 따로 DB로 구축되고 있는 실정이다. 많은 연구와 시스템의 구축으로 교사 인사 업무의 다양한 방면에 도움을 줄 수 있을 것으로 기대된다.

3. 교육을 위한 데이터 마이닝 기법

3.1. 기계학습(Machine Learning)과 교육

기계학습(Machine learning)은 컴퓨터 프로그램을 보다 적응성(Adaptive)있게 해주고 인공지능 프로그램이 극복해야할 지식 습득 문제를 해결할 가능성을 가지고 있기 때문에 인공지능이 싹트기 시작한 시점부터 여러 가지 방식에 대한 연구가 꾸준히 진행되어 오고 있다.

Simon은 프로그램이 하나의 Task를 수행한 후에 그 추론과정에서 얻은 경험을 바탕으로 시스템의 지식을 수정 보완하여 다음 작업에서 처음보다 효과적이고 정확한 문제 해결을 할 수 있는 적응성을 학습이라고 하였다.[5] 학습은 새로운 지식을 습득하거나 새로운 기술을 축적하거나 또는 현재의 지식을 새롭게 구조화하는 과정이라고 볼 수 있다. 인공지능분야에서는 계산적인 방법(Computational Method)으로써 학습의 모형을 표현하며 흔히 기계학습(Machine learning)이란 용어를 사용하여 인지과학 분야에서의 인간학습(Human learning)과 구별한다.

기계학습의 종류는 학습에 사용되는 지식의 종류에 따라 예제에 의한 학습, 경험에 의한 학습, 영역에 의한 학습, 교사에 의한 학습으로 분류된다. 학습방법에 따라 귀납적(Inductive) 학습, 연역적(Deductive) 학습, 사례기반 학습(Case based learning), 유추에 의한 학습, 유전자 알고리즘, 군집화, 신경망 학습, Hybrid 학습 및 Conceptual 학습 등으로 나눈다. 학습의 결과에 따라 지식 습득과 기술 습득으로 나눈다.[3]

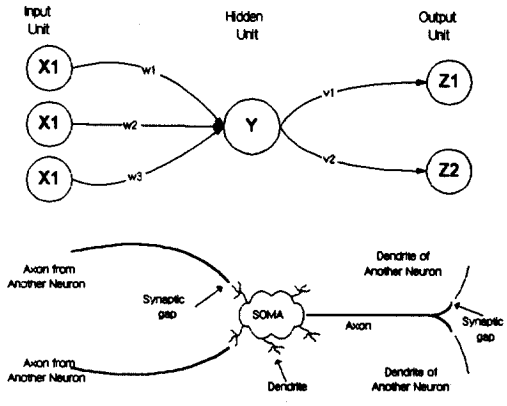
귀납적 학습은 예제들간의 유사성을 기반으로 예제가 암시적으로 나타내는 개념을 추출하는 방법을 사용하므로 유사성 기반 학습(SBL; Similarity Based Learning)이라 하고 ID3, ID5 알고리즘을 이용하여 구현할 수 있다. 연역적 방식은 Domain 이론을 이용하여 하나의 예제로부터 보다 효율적인 지식을 생성한다 이러한 과정은 proof tree를 구축하면서 주어진 예제가 어떻게 개념에 만족될 수 있는가를 설명하기 때문에 설명기반 학습(EBL; Explanation Based Learning)이라고 한다. 사례기반 학습은 일정한 분야에 적용할 수 있는 대표적인 사례들을 기억하고 있다가 새로운 문제를 해결하는 과정에서 유사한 사례를 이용하여 문제를 해결한다.[3]

이러한 기계학습과 교육과의 관계를 살펴보면 학습의 기본이론이 되는 인간의 인지 모델에 기초를 둔다. 교사가 학생을 가르치는 과정과 학생이 기존의 지식을 습득하고, 이러한 지식을 토대로 새로운 지식을 생성하며 문제를 해결하는 과정이 기계학습의 기본 알고리즘과 같다. 이러한 기본 알고리즘은 교사의 교수 모델과 학생의 학습 모델을 모방한 것으로 교육의 교수-학습 과정과 매우 흡사하다. 교사가 학생에게 지식을 효율적으로 전달하기 위해 다양한 방법으로 시도하듯이 기계학습도 인간의 인지 모델을 기초로 하여 다양한 학습 방법을 제시한다. 기계 학습은 컴퓨터의 빠르고 정확한 계산과 인간의 학습, 추론을 통한 병렬 분산 처리 기술을 접목시킨 인공지능 기법으로 대용량의 데이터 처리와 분석, 패턴인식, 음성·문자인식, 지능형 학습 프로그램, 가상 원격학습 등의 교육활동에 다양하게 응용할 수 있다.

3.2. 신경망(Neural Network)과 교육

Neural network은 인간의 두뇌를 모델로 개발하였으며, 프로이드의 정신역학 이론(psycho-dynamic theory)에서 인간의 두뇌를 신경들의 망이라고 주장한데서 비롯된다. 전체를 병렬 분산적으로 운용되는

처리 요소들의 총체로 보고 인간의 신경망과 같은 구조와 역할로 학습을 진행한다.[1]



[그림 5] Neuron과 Neural Network의 구조

상호 연결된 뉴런(처리 요소)에 의해 임의의 입력 N차원 공간을 M차원 공간으로 사상하고, 입력 및 출력 공간과 사상의 특성에 따라 기억, 필터, 변환, 분류, 인식, 최적화 등의 기능 수행한다.

전형적인 신경망은 [그림 5]처럼 노드(node)와 아크(arc)의 집합으로 구성되며 입력 노드는 입력 신호를 받고 아크를 통하여 출력 노드로 전송하여 결과를 출력한다. 신호를 보내게 되면 임의 개수의 중간 계층에 있는 중간 노드들이 다음의 작업을 판단하며, 다양한 연결 형태와 작업을 수행하는 학습전략을 사용한다.

사용단계를 살펴보면 두 단계를 거치게 되는데 첫째, 부호화(encoding)과정으로 신경망이 특정한 작업을 수행하도록 훈련하며, 둘째 단계로 해독(decoding)을 하게 되는데 신경망이 주어진 사례를 분류하거나 예측하며 어떠한 작업도 주어지면 실행할 수 있도록 한다.

신경망의 형태의 형태는 다음 3가지로 나눈다.

인지(perceptron)는 Frank Rosenblatt가 1958년에 제시한 고전적인 모델로 단순 분류 학습 기능이 가능하며, 세 개의 계층으로 된 망으로 구성되었다.

·입력부-광수신기(Photo receptor)

·중간부-연관기(associator)

·출력부-응답기(responder)로 구성되었다.

Back propagation network는 기존의 Perceptron 모형에 은닉 노드를 가지는 여러 개의 중간 계층(은닉 계층)을 도입하였다. 알고리즘을 살펴보면[7]

Step 0. Initial weights and bias.

Set learning rate α. $(0<\alpha\leq 1)$

Step 1. While stopping condition is false, do Step 2~6.

Step 2. For each training pair s:t, do Step 3~5.

Step 3. Set activations of input unit:

$x_i=s_i$.

Step 4. Compute response of output unit:

$y\_in=b+\sum x_i w_i$;

$$\begin{cases} y=1 & \text{if } y\_in>\theta \\ y=0 & \text{if } -\theta\leq y\_in\leq\theta \\ y=-1 & \text{if } y\_in<-\theta \end{cases}$$

Step 5. Update weights and bias if an error occurred for this pattern.

If $y\neq t$,

$w_i(\text{new})=w_i(\text{old})+\alpha t x_i$,

$b(\text{new})=b(\text{old})+\alpha t$.

else

$w_i(\text{new})=w_i(\text{old})$,

$b(\text{new})=b(\text{old})$.

Step 6. Test stopping condition:

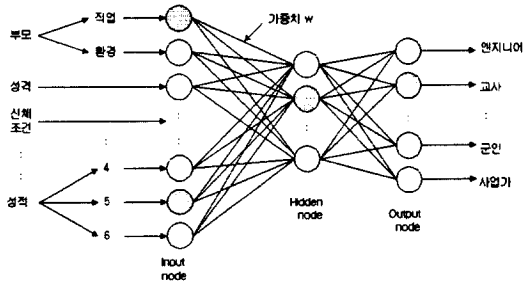
If no weights changed in Step 2, stop; else, continue

이러한 Back propagation network의 특징은 학습이 안정화가 될 때까지 매우 큰 학습 데이터를 필요로 하며, 신경망이 학습을 했을지라도 자신이 배운 것이 무엇인지 제시 못한다는 단점이 있다. 즉 해결과정을 빼고 해답만을 제시하기 때문에 컴퓨터가 학습을 하는 과정에 대해 설명을 할 수가 없다.

세 번째, Kohonen self-organization map 모델은 1981년에 발표된 내용으로 시각지도나 팔다리에 대한 공간적인 지도와 같이 두뇌 속에 존재하는 여러 개의 지도에 대응되는 인공적인 상대물을 표시하는 것으로 뉴런 또는 어떤 단위들이 모여 다시 각각 인접한 작은 개수의 다른 단위들에 연결된다. 초기값의 자가 구성 지도는 각 단위마다 무작위로 할당된 백

터를 가진다. 훈련 단계에서 이 벡터는 주어진 공간을 보다 잘 탐색할 수 있도록 점진적으로 조정되어 학습을 해나갈 수 있다.

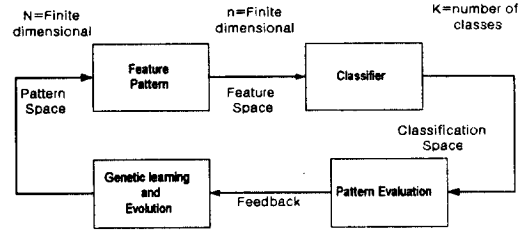
이러한 신경망과 교육의 관계를 살펴보면 신경망을 이용하여 학생의 성격과 환경, 흥미도 등의 자료를 가지고 있는 생활 기록부를 이용하여 학생에게 적합한 직업과 진로를 탐색하여 결정하는 과정을 살펴보면 그림과 같다. 우선, 신경회로망을 위한 전처리 과정을 거쳐야 하는데 이러한 전처리 작업은 자료의 질과 양, 자료의 표현 방법 및 자료 변환에 민감하기 때문에 신중하게 처리해야 한다. 일반적인 자료의 형태는 discrete numeric value, continuous numeric value, categorical (symbolic) value로 표현하며, scaling, normalization, mapping, translation의 방법을 통해서 전처리 과정을 마친다. 이렇게 처리된 학생에 대한 각각의 요인들을 변수로 할당한 후에 이러한 기본자료를 입력 노드에 적용한다. 각 학생의 자료에 대해 신경망의 실제 출력값과 기존에 학습이 되었던 정확한 값에 해당되는 기대 출력값을 비교한다. 그 다음 두 값을 비교하여 차이가 있으면 각 노드와 시냅스에 대한 가중치를 조정하고 정확할 때까지 계속 이 과정을 반복한다. 안정이 되면 결과 값을 출력한다. 이러한 학습 신경망을 구축하기 위해서는 기존의 자료가 매우 정확해야 한다. 졸업을 한 학생의 자료와 진로와 직업이 대용량으로 구축되어야 신뢰도가 높은 결과를 얻을 수 있다. 진로와 직업에 관계된 신경망의 구조를 살펴보면 [그림 6]과 같다.



[그림 6] 신경망을 이용한 진로 교육의 적용 예

3.3. 유전자 알고리즘(Genetic Algorithm)과 교육

Genetic algorithm은 진화적 컴퓨팅(evolution computing)에 기반을 두는데 그에 속하는 내용은 유전자 알고리즘(genetic algorithm), 진화적 프로그래밍(evolutionary programming), 진화 전략(evolution strategy) 등이 이에 속한다. 이러한 것들의 주요 내용은 문제 해결 기법으로 적자 생존(survival of the fittest)의 기법을 사용하며 코딩 문제에 대한 기법으로 DNA모형을 사용한다. 적자생존 모델이 어떻게 유전 정보가 전달되는가에 관한 적절한 이론이 부족하기 때문에 DNA의 나선 구조 모형을 발전시켰는데 유전자 언어로서 네 개의 문자 C, G, A, T로 모두 표현할 수 있는 장점이 있다.[8]

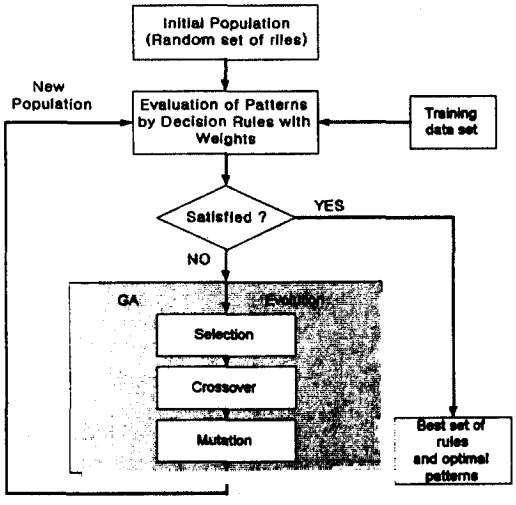


[그림 7] 유전자 알고리즘 Feedback학습 시스템

유전자 알고리즘을 적용하는 절차를 살펴보면 우선, 제기된 문제를 우수하며 훌륭한 코딩 방법으로 제한된 알파벳의 스트링을 설계한다. 그리고 컴퓨터 내에서 해답들이 투쟁 과정에서 서로 결합할 수 있는 인공적인 환경을 만든다. 이 과정에서는 적합 함수(fitness function)라는 성공과 실패를 측정하는 객관적인 기준 제공하며, 다음의 과정으로 예상되는 후보 해답들이 결합될 수 있는 방법을 개발한다. 즉, 부모의 스트링을 잘라 교환한 후에 다시 접합하는 '교잡(cross-over)'연산 방법을 많이 사용한다. 복제에서는 모든 종류의 '돌연변이(mutation)' 연산자가 적용될 수 있다. 마지막으로 잘 분포된 초기 개체들을 만들고 각 세대에서 부실한 해답들은 제거(적자생존)한 후 우수한 해답의 자손이나 변이들로 대체함으로써 컴퓨터에서 '진화(Evolution)'을 할 수 있도록 한다. 그런 후에 성공적인 일련의 해답이 만들어지면 종료한다.[11]

이러한 유전자 알고리즘의 단점은 개체 규모가 과잉으로 발생하며, 탐색과정에서 우연적인 특성이 자

주 발생한다는 문제가 있다. 반면에 견고하기 때문에 해답이 있으면 꼭 발견한다. 응용범위가 넓고 개념이 명료하며 발견된 해답은 기호로 코딩되기 때문에 인간이 알아보기 쉽다는 장점이 있다.



[그림 8] Genetic Algorithm

3.4. Association Rule, CSP와 교육

연관 규칙은 "전체 인원수 중 상위 10%로 학업 성적이 우수하고, 성격이 외향적인 학생들의 90% 이상은 친구관계가 매우 좋다." 와 같이 여러 개의 규칙들이 연관되어 사용하는 것을 연관 규칙이라 하고 이러한 기법을 사용하는 전문가 시스템을 Rule base System이라고 한다.[6] 인간의 뇌는 계산에 있어서 컴퓨터보다 매우 느리고 오류와 기억을 잃기 쉽지만 추론 문제나 여러 가지 규칙이 연결된 복잡한 문제에 있어서는 거대한 병렬 분산 처리 시스템인 신경망이 활동하기 때문에 이러한 문제에 있어서는 컴퓨터보다 매우 정확하고 효율적이다. 주어진 문제의 최적해를 찾기 위해 여러 가지 탐색 기법을 사용하는데 보통 연관되는 규칙이 많아질수록 엄청나게 많은 시간과 비용이 든다. 최적의 해를 탐색하는 기법으로 평가 함수를 이용한 경험적 탐색방법인 휴리스틱(heuristic method)을 많이 사용한다.[2] 최근에 활발하게 연구 중인 제약 만족 문제(CSP, Constrain Satisfaction Problem)가 많이 사용되고 있는데, 이것은 최적의 해 또는 가장 근사한 해를 구하기 위해 여러 가지 제약을 주고 이러한 제약 조건을 만족하는 값을 최적의 해로 구하는 방법이다.[4] CSP의 구성 요소를 살펴보면

Variable $V=\{v_1, v_2, v_3 \dots v_n\}$,
Domain $D=\{d_1, d_2, d_3 \dots d_n\}$
Constrain $C=\{c_1, c_2, c_3 \dots c_n\}$
(단 n은 변수의 개수)

CSP는 변수 V_n에 대하여 각 변수가 들어가야 할 값이 즉, 도메인이 D_n이라고 하면 지수의 비율인 $(V_n!)^{Dn}$의 복잡도를 갖기 때문에 변수나 도메인이 큰 경우 많은 시간이 걸린다. 이러한 문제에서 Forward Checking을 이용하여 제약 조건의 일치성을 검사하고 이에 따라 도메인 여과를 적용하여 이것에 근간을 하여 Backtracking을 할 경우에는 전체 탐색 공간을 줄일 수 있다.[4]

이러한 작업을 줄이기 위해 연관관계는 쉽게 찾아낼 수 있으나 필요 없는 연관 관계도 찾으므로 연관 알고리즘을 실행하기 전에 해당 테이블을 Flattening을 해야 한다. 불필요거나 중복되는 자료를 제거(pruning)해야 한다. 또한 가치 있는 정보를 단지 잡음(noise)에 불과한 정보와 구별하는 것이 매우 어렵기 때문에 의미 있는 연관과 의미 없는 연관을 구별하는 척도가 필요하다. 이러한 방법으로 데이터 베이스에서 많은 사례를 가지는 규칙에 대한 지지율(support)에 대한 표현으로 확신 인자(Confidence Factor)를 부여하여 더욱 정확한 해를 찾을 수 있다.[6]

이러한 연관관계와 제약 만족 문제를 교육과 관련지어 볼 때, 인간의 인지 모델을 토대로 작성되었기 때문에 학생들이 학습시에 문제를 해결해 나가는 방법과 매우 흡사하다. 특히 수학식이나 과학적 추론의 경우처럼 매우 많은 조건이 있는 문제를 해결해야 할 때 연관관계나 CSP를 적용하면 유용하다.

4.교육에의 데이터 마이닝 적용

4.1. 교육용 DB의 데이터 마이닝 적용 부분

대표적인 교육용 DB시스템으로 생활 기록부와 건

강 기록부가 있는데 이를 사용하는 사례를 살펴보면, 우선 생활 기록부를 적용할 수 있다.

적용할 수 있는 분야를 살펴보면 다양한 학생에 대한 성적과 학습 문제에 대한 문제 패턴을 적용하여 효율적인 수준별 학습 방법을 추론할 수 있으며, 범죄 및 폭력에 관한 예방, 학생의 성향을 토대로 한 생활 태도 관찰 및 집단 따돌림 예방, 진단, 건강 예측 및 진단, 아동의 장래 직업 및 진로에 대한 예측, 학부모의 관심 및 성향 등을 추론할 수 있다. 이 중에서 진로에 관계된 내용을 살펴보기로 한다.

진로에 관계된 요인으로 부모와 교사에 대한 간접적인 요인과 학생 자신에 대한 직접적인 요인을 들 수가 있다. 부모의 학력과 직업, 재산 정도, 아동에 대한 관심도, 거주 지역의 분포도, 교사의 태도 및 흥미도, 교사의 기능에 대한 실제 학습 등은 간접적인 요인이며, 아동의 건강상태, 신체 조건, 흥미도, 장래희망, 성적, 지능지수, 특기 및 취미활동, 행동발달 상황(성향)을 토대로 유추해 낼 수가 있다.

이러한 내용을 토대로 학생에게 가장 적합한 진로와 직업을 유추해 내는 작업은 우선 자료를 처리하기 쉽게 전처리(preprocessing) 과정을 거쳐야 한다. 모두 numeric 형태로 변환한 뒤에 cleaning 작업으로 사전에 오류 발견하고, 중복을 제거하는 작업이 필요하다. 특히, 고의로 틀린 정보를 입력한 경우가 있을 수가 있으므로 패턴 인식 알고리즘으로 데이터를 정제하도록 한다. 도메인 일관성이 결여된 경우 즉, 주어진 범위 내에 포함되지 않은 자료가 있는 경우 NULL 값으로 대치하여 Outer join을 한다.[9]

그 다음에 enrichment 작업을 하는데 기존의 레코드 데이터와 조인을 하고 충분한 정보만을 가진 레코드만 선정하여 SQL 연산을 통해 변환한다. 이렇게 변화된 데이터들을 살펴보면 자료들이 패턴인식 알고리즘의 입력으로 사용하기에는 너무 자세한 정보가 되므로 최적의 해를 찾기에 쓸데없는 시간과 비용이 초과되기 때문에 coding 작업을 거친다.

우선, 도메인 값을 축소하여 (예로 학부모의 나이를 10년 정도 간격으로 설정) 시계열 패턴을 발견한다.

-주소를 구역으로

-주민등록 번호를 학생의 나이로

-성적을 5단계의 숫자 등급으로

-학부모의 학력을 숫자 단위로

-'있다/없다'를 0 또는 1로

-해당 날짜를 기준년에서 시작하여 월번호로 변경한다. Flattening-cardinality가 n인 항목은 n개의 이진 항목으로 대치하여 테이블을 축소한다.

이렇게 전처리 과정을 모두 거친 자료들은 실제 데이터 마이닝 기법을 이용하여 유용한 결과들을 도출해 낼 수 있다. 학생의 진로 및 직업을 추론해서 구분(classification)하는 것이기 때문에 이때 사용할 기계학습 기법은 Neural Network와 연관 규칙을 같이 이용하는 것이 좋다. 신경망을 이용하여 직업을 추론하는 내용은 [그림 6]에 설명하였다. 연관 규칙을 이용하여 최적의 해(직업)를 찾기 위해서는 휴리스틱한 방법을 이용한 CSP를 적용하여 학생에 대한 다양한 조건들(변수들)을 다음과 같이 제약 조건으로 표현하여 탐색한다.

변수- 각각의 진로(직업)에 대한 요인들

:Course, FatherDuties, Deed, Body, IQ 등

도메인- 각각의 변수에 대한 값들

:0<Deed<100, 150<Height<200, 80<IQ<150 등

제약조건-각각의 변수가 가지고 있는 제약 조건들

:두 가지 직업을 선택해서는 안 된다.

운동 선수일 경우 Height가 150Cm이하는 안 된다.

등으로 표현하여 제약 조건에 맞는 최적해를 탐색한다. 이에 대한 신뢰도를 확신인자(CF)로 나타낼 수 있다. 이러한 과정은 Visualization 도구를 사용하여 더욱 명확하게 나타낼 수 있다. 이러한 도구로 대표적인 것이 IBM의 Intelligent miner, Integral Solution사의 Clementine, Livingstones 사의 4Thought, SPSS사의 SPSSAnswerTree, SAS사의 Enterprise Miner 등이 있다.

4.2 데이터 마이닝 적용상의 문제점 및 해결 방안

데이터 마이닝을 적용할 경우의 문제점을 크게 나누어 보면 다음과 같이 크게 세가지 경우로 나눌 수 있다.

첫째로, 교육 현장의 DB시스템 구축의 문제가 가장 시급한 부분이다. 현재 종합 생활 기록부의 전산화가 98년부터 일부 학년만 진행 중이며 구축된 시

스템도 안정적이고 체계적이지 않아 DB시스템을 데이터 마이닝 자료로 활용하기에는 아직 미흡한 편이다. 또한 대규모의 자료가 되기에는 전국적인 규모로 볼 때는 충분하지만 학교간의 입력 내용과 방식이 약간씩 차이가 있어 공통적인 데이터의 추출이 어려운 점이 있다. 전국적인 단위로 데이터를 수집하기 위해서 네트워크로 연결되어 있어야 하지만 비용상의 부담과 시설의 어려움으로 이 역시 많은 기간과 비용이 소요되고 있다. 또한 DB 시스템 또한 분산환경의 DB 시스템으로 구축되어야 한다. 현재의 DB 구조가 단일(Stand alone) 사용자용이 대부분이기 때문에 통합의 어려움이 발생된다.

둘째, 자료에 관한 문제로 입력 자료 중 중복되는 자료가 너무 많은 것도 지적된다. 중복되는 자료가 많을 경우 최적의 해를 찾는데 규칙 충돌(Conflict) 현상이 일어 날 수 있으며 탐색소요시간에 과부하를 줄 수가 있다. 또한 중복된 자료를 수정, 제거하는데 많은 비용과 시간이 들게 된다. 입력된 자료의 부적절성과 애매모호한 내용도 큰 문제로 지적할 수 있다. 일정한 형식이 필요한 자료일 경우(예를 들어 날짜 입력) 입력하는 대상이 컴퓨터에 전문적이지 못한 교사들이 있을 수 있기 때문에 오류가 섞인 자료가 있을 수 있고 이 또한 시스템의 성능에 많은 지장을 초래하게 된다. 가장 어려운 부분이 입력 상의 문제로 입력된 내용이 학생의 태도와 성향에 근거하여 정확한가가 의문이다. 교사의 바쁜 업무와 아동에 대한 선입견과 주관적인 생각이 많이 반영될 수 있기 때문에 학생의 다양한 자료가 100%의 신뢰도를 얻을 수 있는지 의문이다.

셋째, 교사의 태도 및 활용, 유지에 관한 문제이다. 우선 보안상의 문제가 있다. 학교 공문서를 취급하는데 있어 개인의 사적인 자료는 지금의 교육현장에서 담임 교사만이 관리할 수 있기 때문에 자료가 공개되는 것을 매우 금기시 하고 있다. 이것은 학습에서의 적용으로 활용되지 못하고 1년이 지난 학생의 귀한 자료가 고문서함에 고이 간직되어 사장되는 결과를 가져오게 되었다. 데이터 마이닝을 적용할 도구가 교육에서는 그렇게 많지 않고 그 모델 또한 매우 빈약하다는 문제가 있다. 도구 또한 비용상으로 매우 만만치 않기 때문에 그 부담도 문제가 된다. 마지막 문제로, 처리된 결과를 활용하는 교사의 태도를 들 수 있다. 처리된 결과에 대해 교사들은 기계의 판단을 믿지 못할뿐더러 처리상의 복잡함으로 인해 이러한 데이터 마이닝 작업을 기피하는 것이다. 또한 그 결과에 대해 학생에 대한 사후 지도 처리과정이 교육 과정상 매우 바쁘기 때문에 사장될 수 있다는 것이다.

이에 대한 해결 방안으로 우선 국가와 정부의 적극적인 지원이다. 교육 예산의 과감한 투자와 H/W와 S/W의 적극적인 지원 등 장기적이고 안정적인 재정의 지원이 절실히 필요하다. 자료의 문제에 있어서는 생활 기록부와 건강 기록부 등 전국적인 단위의 DB 시스템을 정확하고 무결성이 확보되도록 신중하게 고려하여 설계, 작성이 되어야 할 것이다. 교사의 경우 적극적인 연수가 필요하며 21세기 정보화 사회에 대한 안목을 통찰하고, 교수 학습과 담당 업무의 전문성과 일관성을 갖기 위해 부단한 노력을 해야 한다.

5.결론

본 연구에서는 데이터 마이닝에 관한 이론적인 고찰 및 교육에의 적용에 관한 내용을 고찰하였다. 데이터 마이닝의 기법으로 기계 학습에 Neural network, Genetic Algorithm, CSP에 대해 교육과 관계지어 자세히 살펴보았으며, 교육에서 사용되는 대표적인 데이터베이스 시스템으로 종합생활기록과 건강 기록, 성적 자료가 있으며 이를 기계학습 기법과 다양한 데이터 마이닝 기법을 이용하여 유용한 정보를 추출하는 방법과 학습에 이용하는 방법에 대해 살펴보았고, 이러한 데이터 마이닝 기술을 사용하는데의 걸림돌과 자료들을 활용할 때의 문제점과 그에 대한 개선방법을 연구하였다.

차후 연구 과제로 이러한 데이터 마이닝 기법을 사용하여 실제 진로 지도에 관한 시스템과 교육에 관계된 다양한 업무들을 효율적으로 설계 및 구현하여야 하며, 실제 학교 현장에 적용할 수 있도록 대규모의 데이터 베이스 시스템이 구축되기를 기대한다.

참고문헌

[1] 김은주, 이일병 "자료발굴과 신경회로망/기계학습", 한국 경영과학회, SIGDM98 제1호, pp. 62-82, 1988.

[2] 김재희, "인공지능의 기법과 적용", 교학사, pp. 36-245, 1990.

[3] 박영택, 이강로, "ID3 계열의 귀납적 기계학습", 정보과학회지 제13권 제5호, pp.6-19, 1995.

[4] Edward Tsang, "Foundations of Constraint Satisfaction", Academic Press, 1993.

[5] H.S. Simon, "Why should machines learn? In Machine Learning:An Artificial Intelligence Approach", Los Altos, pp. 25-37, 1983.

[6] Joseph Giarranteno, "Expert system :principles and programming", PWS publishing company, pp. 47-55, 1998.

[7] Laurene Fausett, "Fundamentals of Neural Network", Prentice Hall, 1994.

[8] M.Pei, E.D.Goodman, "Pattern Discovery from Data using Genetic Algorithm", Proceeding of the FPAC KDD, pp. 264-276, 1997.

[9] M. Tamer Ozsu, Patrick Valduriez, "Principles of Distributed Database System", Prentice Hall, 1991.

[10] Piatetsky-Shapiro, G.; and Frawley, "Knowledge Discovery in Database". Menlo Park, Calif.: AAAI Press, 1991.

[11] Pieter Adriaans, Dolf Zantinge, "Data Mining", Pentice Hall, 1998.

[12] Ronald J. Brachman and Tej Anand, "The Process of Knowledge Discovery in Data Mining", AAAI Press, pp.37-58, 1996.

[13] Usama M. Fayyad, Gregory Piatetsky-Shapiro, "Advance in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

<URL>

http://info.gte.com/~kdd/

http://www.cs.bham.ac.uk/~anp/TheDataMine.html

http://www.gmd.de/ml-archive

http://www.ics.uci.edu/AI/Machine-Learning.html

http://wwwipd.ira.uka.de/~prechelt/FAQ/neural-net-faq.html

이 철 환

1977년 인천교육대학교 졸업

1987년 연세대학교 교육대학원 졸업(교육학석사)

1988년 한국방송통신대 전자계산학과 졸업(전산학사)

1993년 미국 피츠버그대 정보공학대학원 졸업 (정보공학석사-MSIS)

1994년 미국 피츠버그대학교 사범대학원 교육공학과 컴퓨터교육전공 졸업 (교육학박사-Ed.D)

1989년-1991년 미 피츠버그대학교 사범대학원 교육공학과 연구조교

1991년-1993년 미 펜실바니아주 컴퓨터교육연구소 (ITEC Center) 연구원

1994년 - 현재 인천교육대학교 컴퓨터교육과 교수

연구 분야: 컴퓨터교육, 멀티미디어교육, 웹기반교육

한 선 관

1991년 인천교육대학교(학사)

1995년 인하대학교 전자계산교육과(석사)

1998-1999년 인하대학교 전자계산공학과(박사과정)

1999년~현재 인천교육대학교 컴퓨터교육과 강사

연구 분야: 전문가 시스템, 지능형 교육 시스템, 인공지능, 원격 협력 학습