데이터 마이닝 기법을 이용한 학습 능력 분석 시스템 개발

김 범 은*, 김 덕 희*, 원 유 집**
한양대학교 전자 전기 공학부
candy@hymail.hanyang.ac.kr
thekey95@orgio.net
yjwon@email.hanyang.ac.kr

Application of Data Mining Technique in Characterizing the Scholastic Aptitude of the Students

Beom Eun Kim , Duck Hee Kim, You Jip Won
School of Electronic Engineering, Hanyang University

요 약

데이터 마이닝은 대량의 데이터로부터 데이터 내에 존재하는 관계, 패턴, 규칙등을 찾아내고 모형화 함으로서 유용한 지식을 추출하는 방법이다. 데이터 마이닝을 이용한 이 시스템은 데이터를 비슷한 특성을 가지는 집단으로 분류하여 집단의 특성을 찾아내고 데이터 항목간의 연관성을 유출해 내어 학생들의 적절한 학습지도 영역을 찾아내는데 목적이 있다. 본 논문에서는 개발한 시스템에서 수학 학습 능력에 대한 특성을 도출해 내는 방법을 알아보고, 어떻게 기존의 학원의 역할을 대신할 수 있는지 검증한다.

1. 서론

컴퓨터의 사용이 일반화됨에 따라 학생들의 학습능력을 증진시키기 위한 도구에 컴퓨터 프로그램을 이용하는 경향이 늘어나고 있다. 그러나 이러한 도구들이 아직까지는 데이터베이스화된 문제에서 학생들이 자주 틀리는 부분의 문제를 추출해 반복학습 시킨다거나, 통신을 이용한 On-Line 강의와 같이 임의로 학습의 방법을 바꾸어 학생에게 적용하는 데에 그치고 있으므로 다음과 같은 문제점을 안고 있다.

첫째, 학습의 방법을 바꾸어 학습량을 늘일 경우, 자신이 취약한 부분뿐만 아니라 취약하지 않은 부분까지 접근하므로 효율성이 떨어지며, 학생각자의 특성에 알맞은 학습방법이라고 할 수 없다.

둘째, 자주 틀리는 부분에 대한 반복학습의 경우 어느정도 학생의 취약점을 해결해 줄 수 있으나 근본적으로 학습은 여러 영역이 얽혀 있는 것이므로 관계 있는 다른 영역과 같이 접근해야 하는 근본적인 해결에는 미치지 못하고 있다.

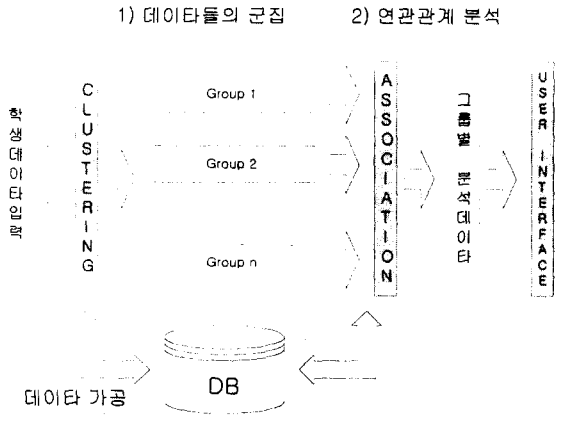
이와 같은 문제를 보완하기 위해서는 먼저 학생의 성적을 분석하여 각 학습요소 데이터간의 관계를 밝혀내어야 할 것이다. 이 문제 해결을 위해 본 프로젝트에서는 데이터 마이닝기법을 이용하여 학습 능력 분석 시스템을 개발하였다.

데이터 마이닝은 대량의 데이터를 분석하여 알려져 있지 않은 정보나 규칙들을 찾아내는 과정이므로 새로운 지식의 발견(knowledge discovery)을 가능하게 한다..(Syllogic, 1996)따라서 여기에서 도출된 정보는 이미 알려져 있고 기대했던 정보뿐만 아니라 전혀 예상하지 못했고 드러나지 않았던 정보까지 포함하며 사용자의 의사결정에 적용시킬 수 있는 의미있는 정보를 말한다.

이 학습 능력 분석시스템은 데이터 마이닝의 여러 기법중 학생들의 수학성적 데이터를 가지고 각 데이터 항목간의 연관성을 분석하기 위해 ASSOCIATION 기법과 한 학생의 데이터만으로는 신뢰성 있는 결과를 얻을 수 없으므로 비슷한 특성을 나타내는 집단끼리 그룹화하는 CLUSTRING 기법을 사용하였으며, 집단의 특성을 파악하고 이를 이용하여 학생들의 적절한 학습지도 영역을 찾아내는 데 목적이 있다. 이 시스템

에서는 결과의 정확성을 위해서 직접 학원으로부터 학생들의 수학성적을 자료로 받아 데이터베이스로 구축하여 결과를 분석하였다.

2. 시스템 구성



[그림 1] 학습 능력 분석 시스템 구성도

[그림 1]에서는 본 연구에서 개발한 시스템의 구성도를 보여주고 있다. 첫번째 Clustering을 이용한 그룹화 부분은 입력 데이터에 따라서 연관있는 데이터를 DB에서 찾아내어 비슷한 특성을 가진 그룹으로 군집화하며 동시에 서로
뚜렷하게 구분되는 군집으로 편성한다. 따라서 한 군집내의 데이터간에는 연관성이 높은 항목들이 포함되어 있다. 두번째 ASSOCIATION을 이용한 연관성 추출부분은 각 군집의 특성을 찾아내고, 데이터 항목들 사이에 연관성을 나타내는 규칙이나 패턴을 도출하여 일정 신뢰수준 이상의 연관성만을 분석에 활용한다. 세번째 USER INTERFACE를 통하여 데이터를 해석하는 부분에서는 입력 데이터가 속한 K번째 군집의 데이터 항목간의 연관성을 보여주고, 다른 군집과의 데이터와 비교하여 현재 입력데이터의 위치를 정확하게 분석해낸다.

이 시스템은 SunOS 5.5.1 sparc Ultra-1 을 사용하였고, DBMS는 Mysql Ver 9.32 Distrib. 3.22.22 for Sun-Solaris 2.5.1, Language는 JAVA version 1.1.8 을 사용하여 개발하였다.

3. 시스템 구현

3.1 Clustering

대량의 데이터를 가지고 그룹의 특성을 찾아내는 이 프로젝트를 효과적으로 실행하기 위해 유사한 데이터들을 몇몇의 집단으로 그룹화하여 각 집단의 성격을 파악함으로서 데이터 전체의 구조에 대한 이해를 돕는 Data Mining 기법인 Clustering을 사용한다. 사용 알고리즘은 대량의 데이터를 군집화하는데 최적의 성능을 가진 K-means Algorithm과 대량의 범주형 데이터(Categorical Data)를 군집화 하는 방법인 K-modes Algorithm[1] 사용하였다.

이 시스템은 학생 개개인의 학습 능력을 정확하게 분석하는 것이 목적이므로 Clustering하고자 하는 목표 데이터 집합(target data set)을 만들고 이를 위해서 입력 받는 데이터의 다양성을 지원한다. 또한 대량의 데이터를 처리하는 속도문제는 데이터 베이스의 데이터로부터 의도하는 목적에 맞게 데이터를 파싱(Parsing)하는 것에 주안점을 두었다.

3.2 Association

Clustering을 사용하여 비슷한 특성을 가진 그룹으로 나누어진 데이터를 가지고 학생의 성적에 영향을 줄 수 있으면서 학습지도에 적용할 수 있는 영역(예를 들면, 수리영역, 추론영역과 같은 문제출제영역과 삼각함수, 미적과 같은 단원출제영역, 각 문제의 난이도등)의 레코드에서 각 데이터간의 상관관계를 찾기 위해 Association기법을 사용하였다. 또한 이 기법을 구현하기 위해 사용된 Apriori Hybrid Algorithm은 기존의 AIS Algorithm이나 SETM Algorithm에 비해 수행속도가 빨라 대용량의 데이터베이스에 적합하다.

Algorithm의 전체적인 자료구조로는 Prune Step(데이터 처리량을 줄이기 위해 데이터를 검색하기 전에 미리 Data Set Candidate중 조건에 맞지 않는 Data Set을 처리하는 과정). 설계하였고, 향상도(Improvement)의 계산을 빠른 수행속도로 수행하기 위해서 Hash Table을 사용하였다. 그리고 Clustering을 적용했더라도 바로 데이터베이스에서 데이터를 가지고 오게 되면 의미 있는 결과를 뽑아낼 수 없으므로 재구성한 데이터를 Input Data로 받도록 설계하였다.

3. 3 Database 구축

데이터베이스는 Exam, ExamGroup, Student, Question, Classification의 테이블로 구성되어 있다. Clustering과 Association의 데이터는 동일 데이터베이스에 연동되어 있으며 Clustering의 결과가 데이터베이스에 저장되어 Association의 입력으로 사용된다.

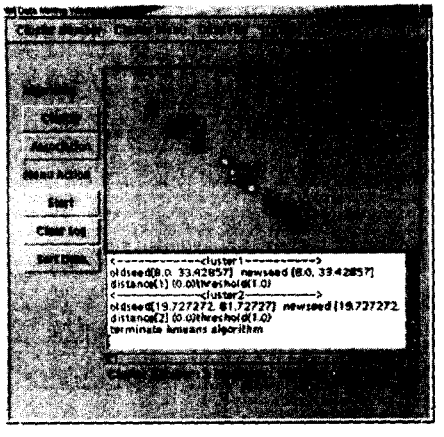
4. 주요기능

결과의 정확성을 위해서 직접 학원으로부터 학생들의 수학

능력 자료를 받아 이 시스템의 학습 능력 평가 자료로 사용하였다. 학생이 수학 능력 자료를 효과적으로 분석하기 위해서 세가지 모드로 Clustering을 한 후 Association을 적용한다.

4.1 Clustering 모드

첫째 학생별 분석 모드는 지금 사용하는 데이터의 경우 학생 개개인에 대한 정보가 불충분하므로 학생과 동일 시험을 본 그룹 내의 학생들의 데이터를 군집화하여 학생이 속한 그룹내 K 번째 군집을 다른 군집과 비교하여 보여준다. 이때 정확한 정보를 얻기 위하여 그룹내 모든 학생에 대해서 각 학생의 시험 총점과 맞은 개수를 두 축으로 잡아서 군집화 한다. 이렇게 하면 맞춘 개수는 적지만 성적이 높은 학생, 맞은 개수는 많지만 성적이 나쁜 학생등 학생들의 다양한 분포를 고려해서 군집화할수 있다.



[그림 2] 학생별 모드의 Clustering 결과 화면

둘째 그룹별 분석 모드는 전체 학생의 데이터를 비슷한 특성을 가지는 K 개의 군집으로 나누고, 각 군집의 특성을 비교, 분석하여 보여준다. 이 경우는 특정 그룹이 다른 그룹과 상이하게 판별되도록 군집화 하여 그룹에 대한 적절한 전략을 세우는 것이 중요하다. 셋째 영역별 분석 모드의 목적은 맞추는 문제 영역(ex. 문제해결력, 추리력, 계산력등) 이 비슷한 학생들끼리 군집화 하는 것이 목적이다.

4.2 Association 모드

Clustering된 결과를 가지고 사용자에게 학습방향을 결정하는데 도움을 줄 수 있도록 연관성을 구해서 보여준다. 그룹별로 군집화 되었을 경우 전체 성적에 관한 군집이므로 각 집단의 영역에 대한 요소(Ex)문제출제 영역, 과목영역)간의 관계에 대한 규칙이나 패턴을 나타내는 Association결과를 그래프화 하여 나타냄으로써 그 그룹에서 좀 더 종속적인 영역에 대한 정보를 알려준다. 또한 어떤 영역의 점수를 올리고 싶은 경우 시너지 효과를 볼 수 있는 다른 영역에 관한 정보와 잠재적으로 취할수 있는 영역에 대한 정보를 알려준다. 그리고 영역별로 군집화 하였을 경우 영역에서의 능력에 따른 전체 성적과의 상관관계를 구함으로써 성적에 따른 집단의 특성을 밝히는데 관해서 정보를 줄 수 있다.

5. 결론 및 향후 연구방향

효율적인 학습을 위한 도구들이 많이 개발되고 있지만 대이터 마이닝을 사용해 학생들의 능력을 분석하여 학습방향의 의사결정에 도움을 주는 시스템은 아직 개발되지 않은 실정이다. 데이터 마이닝을 성적분석에 응용할 경우 각 그룹의 특성과 드러나는 학습영역의 문제점뿐만 아니라 연관성이 있는 다른 영역의 문제점도 발견되어 유용한 정보를 제공하게 되므로 근본적으로 학습능력을 증진시키는 학습방향을 결정하는데 도움을 줄 수 있으리라고 본다.

앞으로의 연구방향은 단순히 그룹에 대한 데이터를 분석하여 정보를 알려주는 데 그치지 않고 성능을 개선하여 분석된 그룹에 관한 정보를 바탕으로 학생 개인이나 그룹이 학습방향을 결정했을 경우 생기는 결과에 대한 예측정보를 제공하여 학습방향을 제시해 줄 수 있는 시스템을 연구하고 설계할 것이다.

6. 참고문헌

[1] Zhexue Huang, "Clustering Large Data Sets with Mixed Numerical and Categorical Values", KDD: techniques and applications, World Scientific, 1997, pp. 21~33

[2] Michael J.A. Berry, Gordon Linoff , Data Mining Techniques for marketing, sales, and customer support, Wiley, 1997

[3] Rakesh Agrawal Ramakrishnan Srikant, Fast Algorithms for Mining Association Rules, VLDB, 1994, pp. 487~499

[4] Data Mining Techniques Data Mining Solution 백서, http://www.sas.com/offices/asiapacific/korea/solution/mining_wp.html

[5] 강 현철외 4 인, 데이터 마이닝 방법론 및 활용, 자유 아카데미, 1999

[6] Aldenderfer, Mark S. Blashfield, Roger K, Cluster Analysis, Sage Publications, 1984