# EasyMiner : 데이터마이닝 시스템 설계 및 구현

최재웅[O] ,류정우, 김종수, 차진호, 손영경, 도영아, 김산성, 이진숙, 김명원
숭실대학교 컴퓨터학과
cyberman@cybersgi.com, mkim@computing.soongsil.ac.kr

# EasyMiner : Design and Implementation of DataMining System

Jae-Woong Choi[O] ,Joung-Woo Ryu, Jong-Su Kim, Jin-Ho Cha, Young-Kyung Son, Young-a Do
San-Sung Kim, Jin-Suk Lee, Myung-Won Kim
Dept. of Computer Science, Soongsil University

## 요 약

정보기술의 발전은 기업들로 하여금 많은 양의 데이터를 기업내부에 축적할 수 있도록 하였지만, 축적된 데이터로부터 기업의 경쟁력을 강화시킬 수 있는 정보를 얻을 수 있는가의 여부는 별개의 문제이다. 즉, 최근 기업들은 최선의 의사결정을 내리는데 필요한 정보 또는 지식을 축적된 데이터로부터 가공해 낼 수 있는가의 여부에 중요한 관심사를 가지고 있다. 데이터마이닝은 바로 이와 같은 요구사항을 충족시키는 새로운 정보기술의 활용방법이다. 본 논문에서는 사용자가 쉽게 데이터마이닝을 접할 수 있게 하기 위해서 데이터마이닝 솔루션인 EasyMiner를 설계하였다. EasyMiner는 데이터베이스에 독립적으로 접근하여, 제공되는 마이닝 기법을 수행할 수 있다. 제공되는 마이닝 기법으로는 분류, 군집화, 연관규칙 그리고 기초통계를 지원하고 있으며 또한 기법들에 의해 생성된 지식들을 사용자에게 쉽게 이해시키기 위해 각 기법의 결과에 대한 가시화를 설계하였다. 본 논문에서는 데이터마이닝 솔루션인 EasyMiner 설계 및 구현에 관하여 제시한다.

## 1. 서론

정보기술의 발전은 기업들로 하여금 많은 양의 데이터를 기업내부에 축적할 수 있도록 하였지만, 축적된 데이터로부터 기업의 경쟁력을 강화시킬 수 있는 정보를 얻을 수 있는가의 여부는 별개의 문제이다. 즉, 기업들이 보유자원의 최적배분을 통하여 최상의 고객만족을 달성하는데 필요한 정보 또는 지식을 축적된 데이터로부터 가공해 낼 수 있는가의 여부가 중요한 관심사가 되었으며, 데이터마이닝은 바로 이와 같은 요구사항을 충족시키는 새로운 정보기술의 활용방법이다.

데이터마이닝이란 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정이다. 여기서 정보는 묵시적이고 잘 알려져 있지는 않지만 잠재적으로 활용가치가 있는 것을 말한다. 다시 말해 기업이 보유하고 있는 일일 거래자료, 고객자료, 상품자료, 마케팅 활동의 피드백 자료와 기타 외부자료를 포함하여 사용 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 실제 경영의 의사결정 등을 위한 정보로 활용하고자 하는 것이다.

이러한 정보를 찾아내는 방법은 어떤 특정 기법과 그 기술 자체만을 의미하는 것이 아니고, 기업의 문제를 이해하고 이러한 문제를 해결하기 위하여 정보기술을 적용하는 포괄적인 과정을 의미한다. 따라서 데이터마이닝을 효율적으로 수행하기 위하여 통계적 기법과 인공지능기법들을 사용하게 된다.

데이터마이닝은 특정문제에 적용하는 기법이 따로 정해져 있지는 않다. 또한 기법이 적용된다고 해서 모든 문제가 해결되는 것도 아니다. 얻고자하는 결과나 데이터의 상태 등에 따라 적용할 수 있는 기법들에 대해 어느 정도의 이해가 수반되면 문제를 해결하는데 있어 보다 효과적이고 적극적인 데이터마이닝을 수행할 수 있을 것이다.

본 시스템에서는 기초적인 통계적 기법과 인공지능 기법을 제공하고 있다. 기초적인 통계적 기법에는 임의적으로 선택된 두 차원의 관련성을 나타내는 산점도와 각 속성에 대한 전체데이터의 빈도수를 나타내는 기초통계가 제공된다. 인공지능 기법으로는 분류(classification), 군집화(clustering), 그리고 연관규칙(association rule)을 제공되며, 각각의 기법에는 다음과 같은 알고리즘들이 제공된다.

본 연구는 한국과학재단 특정기초연구과제
(과제번호 : 98-0102-01-01-3)의 지원을 받았음.

- 분 류 : C4.5 알고리즘
- 군 집 화 : EM, K-means, Cobweb 알고리즘
- 연관규칙 : Apriori 알고리즘

또한 각각의 알고리즘에 의해 생성된 규칙들을 사용자에게 쉽게 이해하고 분석할 수 있도록 데이터 가시화 모듈을 설계하였다.

본 논문의 구성은 다음과 같다. 2장에서는 전체적인 EasyMiner 시스템 구조를 살펴보고, 3장에서는 EasyMiner시스템의 구성요소를 살펴본다. 마지막으로 4장에서 결론을 맺는다.

## 2. EasyMiner 시스템 구조

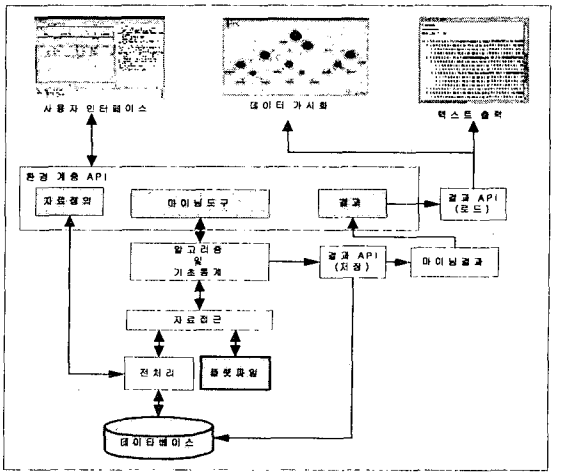


그림 1 . EasyMiner 구조도

EasyMiner 시스템에 대한 전체 구조는 (그림1)과 같다. 자료는 데이터베이스와 플랫파일로부터 입력받을 수 있으며 군집화에 의해 생성된 결과를 다시 데이터베이스에 저장할 수 있으며, 마이닝 결과는 알고리즘에 의해 생성된 모델을 저장한다. 환경 계층 API를 통해 사용자 인터페이스를 사용하여 마이닝 조작 및 마이닝 실행결과를 제어할 수 있다. 결과API(로드)는 마이닝 결과를 사용자에게 그래픽적으로 또는 텍스트형식으로 보여지게 한다.

## 3. EasyMiner 시스템 구성요소

### 3.1 데이터베이스

본 시스템은 데이터베이스와 독립적으로 수행할 수 있다. 즉 현존하고 있는 마이닝 시스템 대부분은 특정 데이터베이스에서만 수행이 되는 반면 본 시스템은 5가지 데이터베이스 (MS-SQL, Oracle, MySQL, DB2, MS-Access)에서 수행 될 수 있다. 특히 MS-Access을 제외한 4개의 데이터베이스에서는 원격 접속 및 전처리가 가능하다. 단, MS-Access는 파일과 같이 단지 로컬에서 수행되어지고 전처리 과정을 수행할 수 없다.

본 시스템에서 제공되고 있는 전처리 기능으로는 SQL문, 레코드 필터, 속성필터, 자료소스조인, 누락값 대체, 누락값 제거 이상 6가지 기능을 제공하고 있다. SQL문은 데이블 생성(CREATE), 테이블 삭제(DROP), 테이블 수정(ALTER), 레코드 삽입(INSERT), 레코드 삭제(DELETE), 레코드 수정(UPDATE), 질의문 작성(SELECT) 이상 7가지의 기본적인 SQL문이 수행되며 테이블명 보여주기(SHOW)은 현재 선택된 데이터베이스에 존재하고 있는 모든 테이블의 이름을 보여주는 명령어로서 본 시스템에서만 제공되는 SQL문이다. 레코드 필터는 SQL함수, 부울연산자, 비교연산자, 산술연산자, 상수, 속성 명을 조합하여 원하는 결과를 얻을 수 있는 과정이다. 속성 필터는 테이블에서 사용자가 원하는 속성만을 선택해서 새로운 테이블을 만든다. 자료소스조인은 두 테이블을 주키(primary key)에 의한 결합으로 새로운 테이블을 만든다. 누락값 제거는 입력 레코드에서 널(null)값이 존재하는 레코드는 제거하는 기능이다. 반면 누락값 대체는 사용자가 원하는 속성에서 대체값을 입력하면 선택된 속성에 존재하는 모든 널값이 대체값으로 바뀌어진다.

### 3.2 알고리즘 및 가시화

본 시스템에서 제공하고 있는 기법과 알고리즘은 표1와 같다.

표 1. EasyMiner에서 제공된 기법 / 알고리즘

| 데이터마이닝 기법 | 알고리즘 |
|---|---|
| 분류 | - C4.5 |
| 군집화 | - K-means<br>- EM Algorithm<br>- Cobweb |
| 연관규칙 | -Apriori |

3.2.1 분류

- C4.5 알고리즘 [1][2][3][6][8]

분류기법에 사용되는 C4.5 알고리즘은 ID3 알고리즘에서 표현할 수 없는 수치적 데이터를 표현할 수 있게 확장한 알고리즘으로 그 결과는 (그림 2)와 같은 트리로 보여준다. EasyMiner에서 제공되는 C4.5 알고리즘에서 수치적 데이터일 경우 이진트리로만 생성한다.

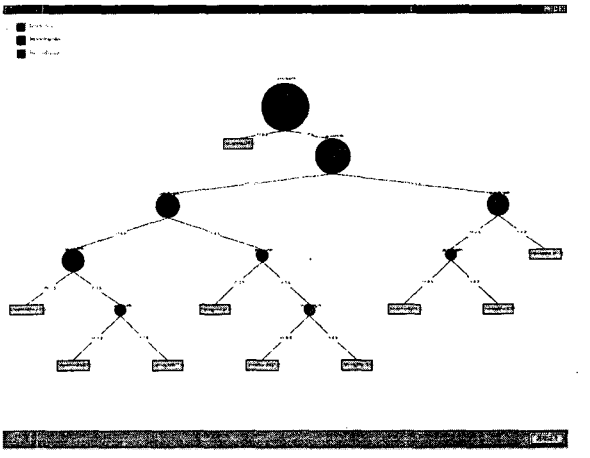

그림 2 분류(C4.5 알고리즘) 가시화

3.2.2 군집화

- EM / K-means알고리즘 [7]

통계적 기법에서 많이 사용되는 EM(Expectation-Maximization) 알고리즘과 가장 간단한 알고리즘으로서 수행속도가 빠른 K-means알고리즘을 제공하고 있다. 알고리즘의 가시화는 (그림3)과 같다. 각각의 클러스터에 포함된 레코드들에 대한 빈도수를 속성별로 나타내고 있으며, EasyMiner에서는 기호적인 속성인 경우에는 이중 파이(pie)차트, 반면 수치적인 속성인 경우에는 히스토그램으로 표현하고 있다. 이중 파이차트에서 외부차트는 전체 데이터에 대한 속성값의 빈도수를 나타내고 내부차트는 생성된 클러스터에 포함된 레코드의 개수를 의미한다. 히스토그램 역시 투명막대는 전체데이터를 색깔막대는 생성된 클러스터를 나타낸다. 또한 알고리즘을 수행하기 전에 속성을 선택할 수 있다. EM, K-means 알고리즘에서는 클러스터 생성시 영향을 미치지는 않지만 생성된 클러스터를 분석하기 위한 속성으로 보조속성을 선택할 수 있다.

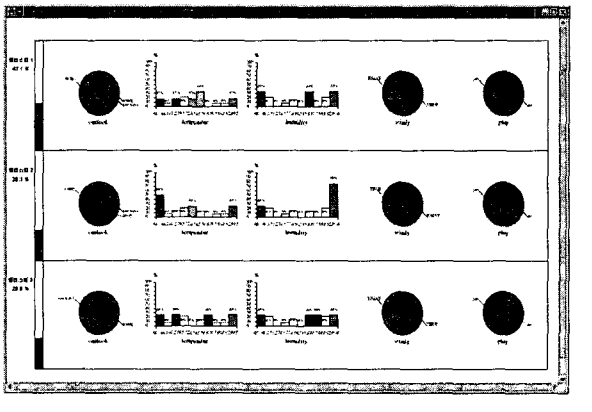

그림 3. EM / K-means 알고리즘의 가시화

- Cobweb 알고리즘 [5]

인공지능 분야에서 텍스트 문서들을 군집화하는데 사용되는 알고리즘으로 앞에서 설명된 두 개의 알고리즘과는 다르게 출력이 (그림4)와 같이 트리로 나타난다.

단말노드가 한 개의 레코드로 구성될 때까지 트리을 생성하여 사용자에게 보여준다. 따라서 Cobweb은 계층적 클러스터링의 한

종류이며 EM 과 K-means와는 다르게 사용자가 사전에 클러스터 개수를 정해 줄 필요가 없다. 즉 생성된 클러스터 개수는 최소 한 개에서 최대 데이터 개수가 될 수 있다. 그러므로 EM과 K-means와 같이 결과를 데이터베이스에 저장할 수 있게 하기 위해 텍스트 출력으로 각 레벨에 포함하고 있는 노드의 개수 즉 클러스터의 개수를 보여주고 사용자는 이 정보를 참고로 하여 클러스터 개수를 결정하는 대신 레벨을 결정하여 데이터베이스에 저장한다.

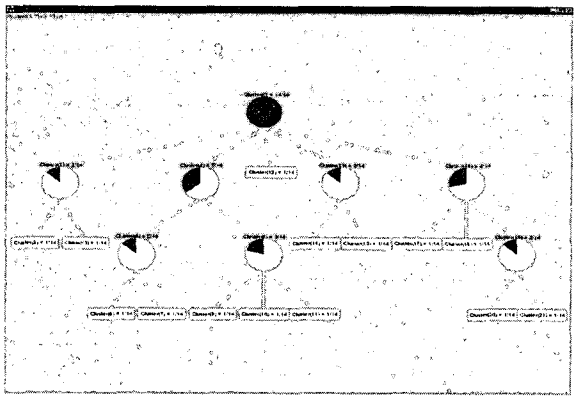

그림 4. Cobweb 알고리즘 가시화

3.2.3 연관규칙 [4]

- Apriori 알고리즘

Apriori 알고리즘은 생성된 연관규칙이 전체 항목에서 차지하는 비율을 말하는 지지도(support degree)와 연관규칙의 강도를 의미하여 전제부를 만족하는 항목이 결론부까지를 만족하는 '비율인 신뢰도(confidence degree)를 이용하여 연관규칙을 추출한 다음 그 결과를 (그림 6)과 같이 사용자에게 보여준다.

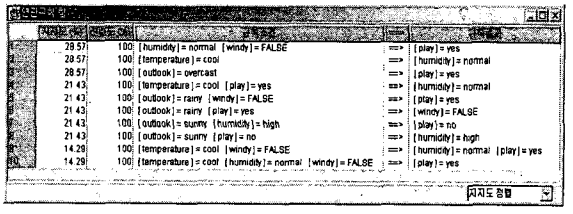


그림 5. 연관규칙 가시화

## 4. 결론 및 향후 연구과제

현존한 데이터마이닝 시스템에서는 특정한 데이터베이스에서만 수행되는 반면 본 시스템에서는 JDBC드라이버를 사용하여 5개의 데이터베이스에서 수행될 수 있으며, MS-Access을 제외한 4개의 데이터베이스는 원격으로 수행될 수 있도록 설계 및 구현하였다. 또한 사용자가 마이닝한 결과를 쉽게 이해하고 분석할 수 있게 하기 위해서 각각의 알고리즘의 특성에 따라 가시화를 설계 및 구현하였다.

향후 계획으로는 데이터마이닝에서 생성할 수 있는 지식의 종류의 확장과 보다 효율적인 데이터 마이닝 기법들의 알고리즘 개발이 필요하며, 본 시스템을 Web상에서 마이닝을 할 수 있도록 확장되어야 할 것이다.

## 참고 문헌

[1] Frank, E. and I. Witten. 1998. Generating accurate rule sets without global optimization. In Shavlik, J. , editor, Proc Fifteenth International Conference on Machine Learning, Madison, WI. San Francisco: Morgan Kaufmann, pp 144-151.-.1999. Making better use of global discretization. In Brotko, I. and S. DzeBled, Slovenia. San Francisco: Morgan Koufmann, pp.115-123.

[2] Wang, Y. and I.H. Witten. 1997. Induction of model trees for predicting continuous classes. In van Someran, M. and G.Widmar, editors. Proc.of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague, pp.128-137.

[3] Atkeson, C. G., S. A. Schall, and A. W. Moore. 1997. Locally weighted Learning. AI Review 11, pp. 11-71.

[4] Chen, M. S., J. Jan, and P.S. Yu. 1996. Data mining: An overView from a database perspective. IEEE Trans. Knowledge and Data Engineering 8(6), pp 866-883.

[5] Fisher, D. 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning2(2), pp.139-172

[6] Gaines, B. R., and P. Compton. 1995. Induction of ripple-down rules applied to modeling large data bases. Journal of Intelligent Information Systems 5, pp. 211-228.

[7] Alfred O. Hero and Jeffrey A. Fessler, 1995, Convergence in norm for alternating expectation-maximization(EM) type algorithms The University of Michigan, Statistica 5, pp 41-54.

[8] 1993. Quilan, J. R. C4.5:Programs for machine learning. San Francisco: Morgan Kaufmann.