Journal of the Korean Data & Information Science Society
2009, **20**(2), 425–433

한국데이터정보과학회지

피에이치피와 웨카를 이용한 데이터마이닝 도구의 설계 및 구현

유영재[1] · 박희창[2]

[1]창원대학교 생명정보학과 · [2]창원대학교 통계학과

접수 2009년 1월 24일, 수정 2009년 3월 18일, 게재확정 2009년 3월 23일

요 약

데이터마이닝은 방대한 양의 데이터 속에서 유용한 정보를 찾아내는 과정이며, 이를 위해 데이터마이닝 도구가 필요하다. 데이터마이닝 도구 또는 솔루션은 E-Miner, Clementine, WEKA, R 등 상당히 많은 종류가 있으나 대부분의 데이터마이닝 도구는 다양성과 범용성에 초점을 맞추어 개발되어 사용 편의성과 분석 자동화에 대해서는 소홀한 실정이어서 비전문가가 사용하기 어려운 경우가 대부분이다. 본 논문에서는 피에이치피와 웨카를 이용하여 인터넷 환경에서 데이터마이닝 기법을 실행하고, 생성된 분석결과를 보다 쉽게 해석할 수 있도록 개선하여 일반 사용자도 쉽게 사용할 수 있는 시스템을 설계하고 구현하고자 한다. 본 논문에서 구현하는 데이터마이닝 기법은 가장 많이 이용되고 있는 연관성 규칙의 Apriori 알고리즘, 군집분석의 K-평균 알고리즘, 의사결정나무의 J48 알고리즘 등이다.

주요용어: 군집분석, 데이터마이닝 도구, 연관성규칙, 의사결정나무.

1. 서론

데이터마이닝(data mining)은 대량의 데이터로부터 정보를 추출하고 이를 바탕으로 의사결정에 이용하는 것을 의미한다. 사전적 의미로 데이터에서 채굴한다는 의미로 자료에서 가치 있는 것을 캐내는 작업을 말한다 (Han과 Kamber, 2001). 데이터마이닝은 정보 기술의 발달과 더불어 발전되어 왔다. 특히 데이터베이스 기술의 발달과 인공지능의 전문가 시스템과 기계학습 등이 데이터마이닝을 발달시키는데 주요한 요인을 제공하였다. 데이터마이닝은 현재 광범위한 영역에서 활용되고 있으며, 기업에서 뿐만 아니라, 공공기관, 생명공학 등 보다 복잡한 정보 분석이 요구되는 많은 분야에서 데이터마이닝을 활용하고 있다 (이창호 등, 2000).

데이터마이닝 도구 또는 솔루션은 SAS E-Miner, SPSS Clementine, WEKA, R 등 상당히 많은 종류가 있다. 이들을 활용한 대표적인 연구로는 Kim (2003)이 데이터마이닝에서의 분류방법에 관한 연구를 위해 E-Minner를 이용한 바 있으며, Kang 등 (2003)은 보험 CRM에서 데이터마이닝 기법의 응용에 대해 연구하기 위해 E-Miner를 이용하였다. Park과 Cho (2005a, 2005b)는 연관성 규칙과 의사결정나무를 이용하여 사회지표조사자료와 폐기물 데이터를 분석하기 위해 Clementine을 활용한 바 있다. 또한 박인우와 권재기 (2007)는 대학의 성공적인 ERP 구축을 위한 대학 특성의 유형을 분석하기 위해 웨카를 이용하였으며, 김성수 등 (2005)은 R을 이용하여 회귀분석과 실험계획법에 대한 시스템을 구축

[1] (641-773) 경남 창원시 사림동 9번지, 창원대학교 대학원 생명정보학과.
[2] 교신저자: (641-773) 경남 창원시 사림동 9번지, 창원대학교 통계학과, 교수.
E-mail: hcpark@changwon.ac.kr

한 바 있다. 그러나 이러한 데이터마이닝 도구들은 기법의 다양성과 범용성에 초점을 맞추어 개발되기 때문에 개별 데이터마이닝 기법들에 대한 사용 편의성과 자동화에 대해서는 소홀한 실정이어서 비전문가가 사용하기 어려운 경우가 대부분이다.

따라서 인터넷이 가능한 환경이면 어디서나 데이터마이닝 분석을 실행하고 생성된 분석결과를 보다 쉽게 해석할 수 있도록 지원하여 일반 사용자도 쉽게 사용할 수 있도록 기존 데이터마이닝 도구들의 단점인 사용 편의성을 개선한 시스템이 필요하다. 이러한 시스템의 개발을 위해서는 범용적으로 이용되는 피에이치피와 웨카를 응용하는 것이 바람직하다고 생각된다. 시스템의 사용자 인터페이스와 웹 프로그램 개발에 사용되는 피에이치피는 강력한 성능과 편리하게 사용할 수 있는 스크립트 언어로써 1994년 Lerdorf에 의해 만들어졌으며, 명령 해석 엔진과 도구들로 구성되어 있다. 2001년 NETCRAFT의 통계 자료에 의하면 피에이치피는 전 세계의 300,000대 이상의 웹 서버에서 사용 중이며, 도메인에서 사용 중인 피에이치피까지 포함한다면 두 배 이상이 될 것이라고 추정하고 있다 (정인근 등, 2002). 그리고 시스템의 분석 모듈로 사용되는 웨카는 뉴질랜드 와이카토 대학교의 Witten 교수팀이 개발한 기계학습 알고리즘이다 (Holmes 등, 1994). 웨카는 Java 언어로 구축된 오픈 소스 소프트웨어 도구(GNU GPL)로써 대학 또는 대학원의 교과 학습이나 그 외 연구 목적에서 누구나 쉽게 사용할 수 있어 새로운 기계 학습 알고리즘 개발에 활용될 수 있다. 또한 웨카 내부에는 일반적으로 데이터마이닝 분야에서 필요로 하는 기본적인 알고리즘의 대부분이 구현되어 있는 동시에 데이터 분석을 위한 사용자 도구도 포함되어 있다. 뿐만 아니라 웨카는 고가의 상용 도구에 비해서도 뒤지지 않는 기능을 제공하면서도 어느 누구나 무료로 사용이 가능하여 많은 분야에서 활용되고 있다. 하지만 분석결과의 해석을 도와줄 수 있는 사용자 도구가 부족한 것이 단점이라고 할 수 있다.

본 논문에서는 피에이치피와 웨카를 이용하여 웹 환경 하에서 데이터마이닝을 실행하는 동시에 웨카의 단점을 보완하여 분석결과를 용이하게 해석할 수 있는 시스템을 설계하고 구현하고자 한다. 2절에서는 피에이치피와 웨카를 이용하여 웹 환경 하에서 데이터마이닝을 실행하고 생성된 분석결과를 쉽게 해석할 수 있는 시스템의 설계에 대해 기술하며, 3절에서는 시스템의 구현 결과에 대해 기술한 후 4절에서 결론 및 향후과제를 다루고자 한다.

2. 웹 기반 데이터마이닝 도구의 설계

본 논문에서 설계하고 구현하고자 하는 시스템은 웹 기반의 데이터마이닝 도구로써 도구의 설치과정 없이 인터넷이 가능한 환경이면 어디서나 데이터마이닝 기법을 실행하고 생성된 분석결과를 보다 쉽게 해석할 수 있도록 지원하는 시스템이다. 시스템의 구성은 크게 데이터 관리, 분석설정, 분석, 분석결과 관리로 구성되어 있으며, 시스템의 구조는 다음의 그림 2.1과 같다.

시스템은 크게 사용자 인터페이스를 포함하는 클라이언트, 클라이언트의 접속 및 명령을 수행하는 웹서버, 웹서버와 웨카를 연결하는 bridge, 데이터를 분석하는 분석 모듈로 구분할 수 있다. 클라이언트는 사용자 인터페이스에 접속할 수 있는 웹 브라우저이며 웹서버는 클라이언트가 시스템에 접속할 수 있게 해주며 전송된 명령을 해석하여 분석 모듈을 제어하거나 분석결과를 클라이언트에게 전달하는 역할을 한다.

본 절에서 구현하고자 하는 웹 환경 하에서 데이터마이닝 알고리즘은 그림 2.1의 시스템 구조도에서 나타나는 바와 같이 연관성 규칙의 Apriori, 군집 분석의 K-평균, 의사결정나무 기법의 J48 등이다. 연관성 규칙은 대용량 데이터베이스에서 각 항목들 간의 관련성을 찾아내는 기법으로 Agrawal 등 (1993)에 의해 처음 소개되었다. Apriori 알고리즘은 후보 항목 집합을 구성하고, 발생 빈도수를 계산하고 난 후에 사용자가 정의한 최소 지지도를 기초로 빈발 항목 집합들을 결정한다.

군집분석은 다양한 특성을 지닌 관찰대상을 유사성을 바탕으로 동질적인 집단으로 분류하는데 쓰이는

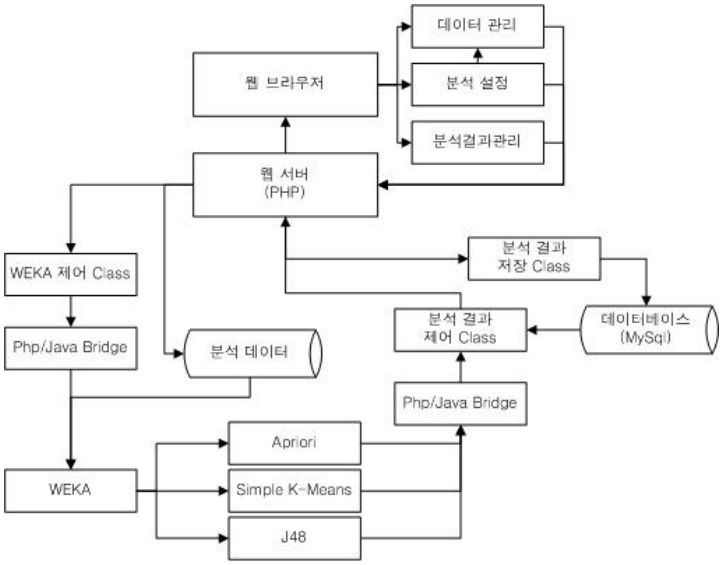

그림 2.1 시스템 구조도

기법이다. 군집분석의 시초라고 할 수 있는 K-평균은 MaxQueen (1967)에 의해 처음 소개된 알고리즘이며, 데이터들을 k개의 군집으로 임의로 분할을 하여 군집의 평균을 대표값으로 분할해 나가는 방법으로 데이터들을 유사성을 바탕으로 재배치를 하는 방법이다.

의사결정나무는 의사결정규칙을 나무구조형태로 도표화하여 관심의 대상이 되는 집단을 여러 개의 소집단으로 분류하거나 예측을 수행하는 분석기법이다. 대표적인 의사결정나무 알고리즘에는 Hartigan (1975)이 제안한 Chaid, Breiman 등 (1984)이 제안한 Cart와 Quinlan (1992)에 의해 제안된 C4.5 등이 있다.

C5.0 알고리즘은 다지 분리를 수행하는 알고리즘으로 엔트로피(entropy)를 불확실성의 측도로 이용하여 예측변수의 기준으로 사용하는데, J48은 C4.5를 웨카에서 확장해서 구현한 알고리즘이다.

본 논문에서 구현하고자 하는 데이터마이닝 시스템의 흐름도는 다음의 그림 2.2와 같다.

분석을 진행하기 위해서는 분석할 데이터를 선택해야 한다. 새로운 데이터를 분석할 경우는 데이터 등록을 통해 서버에 저장하여야 한다. 데이터를 선택하고 나면 시스템은 데이터를 분석하여 변수들을 나열해 주며 사용자는 변수를 선택하거나 제외하여 분석에 포함하거나 제외시킬 수 있으며, 각 변수의 속성을 확인할 수 있다. 변수의 선택과 함께 분석 알고리즘을 선택하고 해당 알고리즘의 속성을 설정할 수 있다. 설정된 변수와 속성은 서버를 통해 분석 모듈에게 전달되며, 분석 모듈은 분석을 시행하여 결과를 분석결과 전처리기에 전달한다. 전달된 분석결과는 분석결과 전처리기에 의해 클라이언트에서 사용 가능하도록 변형된다. 변형된 분석결과는 다시 웹서버를 통해 클라이언트의 분석결과 후처리기에 전달된다. 분석결과 후처리기는 분석결과를 해석하여 브라우저에 전달하여 사용자가 확인할 수 있도록 해준다. 분석결과 후처리기는 이러한 해석과 전달 기능 외에 사용자의 요구에 의해 결과를 정렬하거나 입력한 조건에 맞추어 결과를 검색하는 기능을 함께 수행한다. 또한 검색결과는 검색결과 저장기능을 통해 저장해 두었다가 언제든 다시 확인할 수 있다. 이렇게 저장된 분석결과를 확인하는 경우 사용자는 저

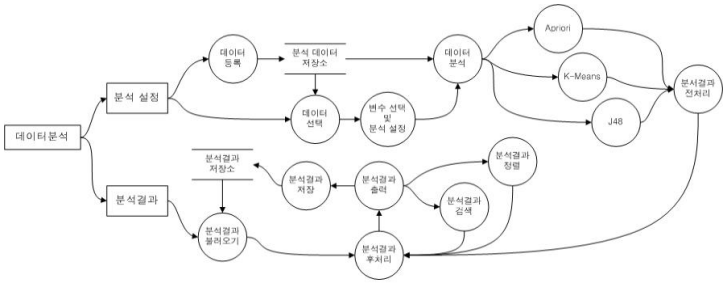

그림 2.2 시스템 흐름도

장된 분석결과 불러오기 기능을 통해 이전에 진행되었던 분석결과를 언제든지 확인이 가능하다. 사용자는 분석 데이터, 분석설정, 저장일 등을 확인하여 원하는 분석결과를 선택할 수 있으며, 선택된 분석결과는 분석결과 후처리기에 전달되어 이전에 분석한 후에 확인했던 화면과 동일한 화면에서 확인할 수 있다.

3. 웹 기반 데이터마이닝 도구의 구현

클라이언트의 웹 접속을 가능하게 해 주고 분석 및 데이터의 요청에 대한 처리 및 결과 반환, 데이터 및 분석결과를 저장하는 역할을 하는 서버에 사용된 운영체제는 Linux이다. 웹서버는 Apache, 웹 프로그래밍 언어는 피에이치피, 데이터베이스는 MySQL이 사용되었으며, 분석 모듈은 웨카 3.4.1이 사용되었다. 또한 Java로 개발된 웨카와 피에이치피를 연동하기 위해 피에이치피/java Bridge를 사용하였다. 클라이언트는 사용자 인터페이스로써의 역할과 분석결과를 제어하여 원하는 결과를 쉽게 찾을 수 있도록 하는 역할을 한다. 이러한 클라이언트의 개발은 웹 브라우저를 기반으로 하기 위해 XHTML 1.0, JavaScript, CSS를 사용하여 개발하였으며, 대부분의 기능은 JavaScript에 의해 제어된다. 데이터의 송수신은 AJAX (asynchronous javascript and XML)를 통해 이루어진다.

실제 구현된 데이터마이닝 도구의 구현 화면 중 분석하고자 하는 데이터를 선택하는 화면은 다음의 그림 3.1과 같다. 상단의 파일 선택을 누르면 아래에 저장된 데이터의 목록이 출력되고 원하는 데이터를 누르면 선택된다. 새로운 파일을 분석하고 싶을 때는 파일업로드를 눌러 원하는 파일을 선택하면 서버에 저장된다.

데이터 선택 후 분석하고자 하는 변수를 선택하는 화면은 다음의 그림 3.2와 같다. 데이터를 선택하면 분석데이터 항목에 해당 데이터의 이름이 나타나고 그 아래에 변수들이 나열된다. 변수는 기본적으로 선택된 상태이며 분석에서 제외하고자 하는 경우 체크를 해제할 수 있다. 또한 분석 알고리즘에 따라 사용할 수 없는 자료형의 변수는 비 활성화되며 선택이나 해제할 수 없다. 변수를 선택하면 우측에 해당 변수에 대한 기본 정보가 출력된다.

알고리즘별 분석설정 화면은 다음의 그림 3.3과 같다. 각 알고리즘별로 분석에 대한 설정을 할 수 있다.

알고리즘별 분석결과 화면은 다음의 그림 3.4, 그림 3.5, 그리고 그림 3.6과 같다.

그림 3.4는 weather.nominal.arff 데이터에 대해 10개의 연관성 규칙을 생성한 결과 화면이다. 상단에 데이터명과 알고리즘, 분석 설정이 출력되며 그 아래 검색 입력 창과 연관성 규칙이 표 형식으로 출

그림 3.1 데이터 선택화면

그림 3.2 변수 선택화면

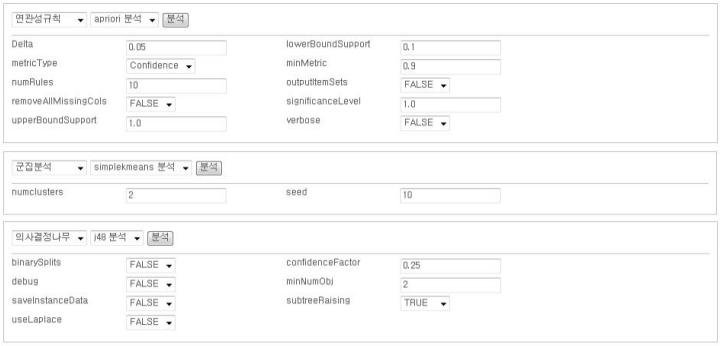

그림 3.3 분석 설정 화면

력된다. 연관성 규칙의 상단에 Antecedent Consequent 등과 같은 레이블을 누르면 해당 항목을 기준으로 오름차순과 내림차순으로 번갈아 정렬순서를 변경해서 볼 수 있다. 또한 상단의 검색 입력창에 원하는 조건을 입력하고 분석결과 검색을 선택하여 원하는 연관성 규칙만 찾아서 볼 수 있다.

그림 3.5는 iris.arff 데이터를 2개의 군집으로 분류한 결과화면이다. 연관성 규칙의 결과 화면과 동일하게 분석에 대한 정보와 검색 입력창, 분석결과창이 출력된다. 군집 분석도 분석결과에 대한 정렬기능과 검색 기능을 제공한다.

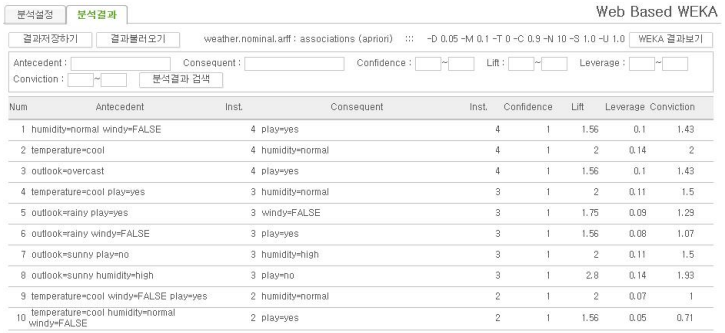

그림 3.4 Apriori 결과 화면

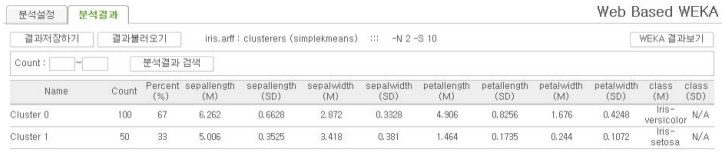

그림 3.5 Simple K-평균 결과 화면

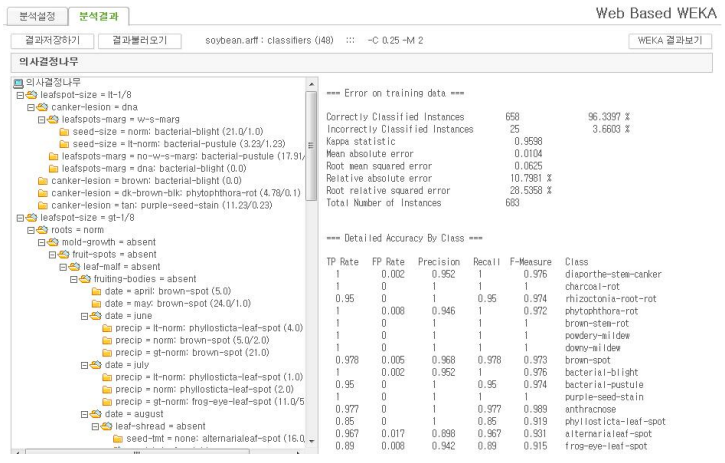

그림 3.6 J48 결과 화면

그림 3.6은 soybean.arff 데이터를 J48 알고리즘을 이용해 의사결정나무를 생성한 결과 화면이다. 상단에 데이터명과 알고리즘, 분석 설정이 출력되며 그 아래 경로 출력창이 있다. 그 아래쪽의 좌측에는

의사결정나무가 노출되며 우측에 기타 정보가 출력된다. 의사결정나무의 노드를 선택하면 경로 출력창에 해당 노드에 대한 경로를 출력하여 결과 확인을 도와주며, 자식 노드에 대한 보이기/숨기기 기능을 통해 원하는 노드만 보여 줄 수 있다. 분석결과들은 결과저장하기를 통해 서버에 저장해 두고 언제든지 다시 확인할 수 있다. 이렇게 저장된 분석결과를 불러오는 화면은 다음의 그림 3.7과 같다.

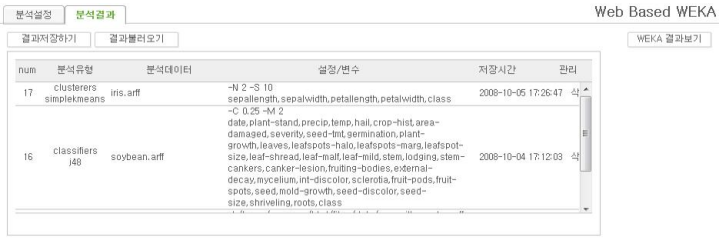

그림 3.7 분석결과 선택 화면

결과 불러오기를 누르면 저장된 분석결과들을 출력해주고 원하는 분석결과를 선택하면 이전의 분석결과 화면과 동일한 분석결과를 제공한다. 분석결과 목록에는 분석방법, 분석 알고리즘, 분석데이터, 분석설정, 분석변수, 저장시간을 출력하여 원하는 분석결과를 찾을 수 있도록 해준다. 또한 상단 우측의 웨카 결과보기를 통해 분석결과에 대한 전/후처리 과정을 거치지 않은 웨카의 분석결과를 확인할 수 있다.

4. 결론

본 논문에서 설계하고 구현한 피에이치피와 웨카를 이용한 웹 기반의 데이터마이닝 도구는 설치과정 없이 인터넷이 가능한 환경에서 데이터마이닝 분석을 실행하고 생성된 분석결과에 대한 정렬, 검색 등의 기능을 통해 일반 사용자도 보다 쉽게 사용할 수 있는 시스템을 설계하고 구현하였다.

데이터마이닝 도구 중 하나인 웨카는 훌륭한 데이터마이닝 도구이나 분석의 전처리 과정과 분석의 설정이 다소 복잡하고 분석결과가 대부분 문장 형태로 제공되어 연관성 규칙과 같이 다량의 분석결과가 나타날 경우 분석자가 결과를 분석하는데 어려움이 있다. 또한 분석결과 저장 또한 문장 형태의 결과 그대로 파일로 저장해 둘 수밖에 없기 때문에 분석결과를 다른 분석에 활용하기 어렵다. 이에 본 연구에서는 데이터마이닝의 비전문가도 쉽게 데이터마이닝 분석을 진행할 수 있도록 웨카의 복잡한 설정 과정을 단순화하고 웹 기반으로 개발하여 설치 과정 없이 언제 어디서나 분석을 진행할 수 있도록 하였다. 연관성 규칙 중 Apriori, 군집분석 중 K-평균, 의사결정나무 중 J48 등의 3가지 알고리즘을 사용하여 데이터마이닝 분석결과를 제공하며 분석결과 해석을 돕기 위해 연관성 규칙의 결과를 변수별로 오름차순과 내림차순으로 번갈아 정렬순서를 변경해서 볼 수 있는 정렬 기능과 원하는 조건으로 분석결과를 찾아볼 수 있는 검색 기능을 추가하였다.

본 연구에서는 알고리즘의 다양성보다는 웨카를 활용한 웹기반 데이터마이닝 도구의 개발 가능성에 대해 초점을 맞추어 위에서 기술한 3가지 알고리즘만 적용하였다. 그러나 향후에는 더욱 더 다양한 알고리즘의 접목과 쉽고 편리한 사용자 인터페이스에 대한 연구를 통해 보다 나은 데이터마이닝 도구의 개발이 필요할 것이다.

참고문헌

김성수, 박희진, 조영훈, 오진호 (2005). R을 이용한 회귀분석과 실험계획법 시스템 구축. <한국통계학회 2005 추계 학술발표회 논문집>, 5-11.

박인우, 권재기 (2007). 대학의 성공적인 ERP 구축을 위한 대학특성 유형분석. <교육문제연구>, **29**, 73-101.

이창호, 이남근, 이승희, 이병엽, 김주용 (2000). 시나리오 기반의 데이터마이닝 도구 XM-Tool/Miner 설계 및 구현. <한국지능정보시스템학회 2000년 학술대회논문집>, **2,** 307-314.

정인근, 이명무, 김용진 (2002). Perl/CGI와 피에이치피의 비교를 통한 웹 어플리케이션 개발성과에 미치는 영향에 관한 연구. <한국경영과학회 2002 춘계학술대회논문집>, 58-64.

Agrawal, R., Imielinski R. and Swami, A. (1993). Mining assocation rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 207-216.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and regression trees*, Wadsworth International Group, Belmont, California.

Han J. and Kamber M. (2001). *Data mining: Concepts and techniques*, Morgan Kaufmann, San Francisco.

Hartigan, J. A. (1975). *Clustering algorithms*, John Wiley & Sons, Inc, New York.

Holmes, G., Donkin, A. and Witten, I. (1994). WEKA: A machine learning workbench. *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, 357-361.

Kang, H. G., Kim, K. K., Kang, C. W., Choi, S. B. and Cho, S. K. (2003). Applied study on data mining technique in insurance CRM. *The Journal of Korean Data Analysis Society,* **5**, 101-112.

Kim, K. K. (2003). A study on classification methods in data mining. *The Journal of Korean Data Analysis Society*, **5**, 101-112.

MaxQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Park, H. C. and Cho, K. W. (2005a). Waste database analysis joined with local information using association rules. *The Journal of Korean Data Analysis Society*, **7**, 763-772.

Park, H. C. and Cho, K. W. (2005b). Social indicator survey data analysis using decision tree. *The Journal of Korean Data Analysis Society*, **7**, 773-783.

Quinlan, J. (1992). C4.5: *Programs for machine learning*, Morgan Kaufmann, San Francisco.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 425–433

한국데이터정보과학회지

Design and implementation of data mining tool using PHP and WEKA

Young Jae You[1] · Hee Chang Park[2]

[1]Department of Bioinformatics, Changwon National University
[2]Department of Statistics, Changwon National University

Received 24 January 2009, revised 18 March 2009, accepted 23 March 2009

Abstract

Data mining is the method to find useful information for large amounts of data in database. It is used to find hidden knowledge by massive data, unexpectedly pattern, relation to new rule. We need a data mining tool to explore a lot of information. There are many data mining tools or solutions; E-Miner, Clementine, WEKA, and R. Almost of them are were focused on diversity and general purpose, and they are not useful for laymen. In this paper we design and implement a web-based data mining tool using PHP and WEKA. This system is easy to interpret results and so general users are able to handle. We implement Apriori algorithm of association rule, K-means algorithm of cluster analysis, and J48 algorithm of decision tree.

Keywords: Association rule, cluster analysis, data mining tool, decision tree.

[1] Department of Bioinformatics, Changwon National University, Changwon 641-773, Korea.
[2] Corresponding author: Professor, Department of Statistics, Changwon National University, Changwon 641-773, Korea. E-mail: hcpark@changwon.ac.kr