공개 소스시스템을 이용한 효과적인 마이닝 전략

An Effective Mining Strategy Using Open Source System

전성해[1], 이승주[1], 오경환[2]

[1] 청주대학교 바이오정보통계학과
E-mail: {shjun, access}@cju.ac..kr
[2] 서강대학교 컴퓨터공학과
E-mail: kwoh@sogang.ac..kr

요 약

CRM, Bioinformatics 등 데이터 마이닝이 적용되는 분야에서 데이터분석에 주로 사용되는 도구는 고가의 마이닝 패키지들이다. SAS사의 Enterprise Miner, SPSS사의 Clementine 등이 대표적이다. 그러나 이 패키지들은 방대한 분석시스템으로 이루어져 있기 때문에 배우기가 어렵고 또한 고급의 컴퓨터 리소스들을 요구한다. 따라서 중소규모의 업체나 연구조직에서는 이와 같은 고성능의 마이닝 패키지를 사용하는데 상당한 어려움이 있다. 본 논문에서는 공개 소스시스템에서 제공되는 패키지를 이용하여 중소규모의 마이닝 프로젝트를 위한 효과적인 전략을 제안한다. 본 논문의 제안전략에 의해 비용 절감과 동시에 수준 높은 마이닝 결과를 제공받을 수 있게 된다.

1. 서 론

자동화된 데이터 수집도구와 눈부시게 발전하는 데이터베이스기술로 인하여 대단히 많은 데이터가 데이터베이스 관리시스템(DBMS: database management system) 혹은 데이터웨어 하우스(DW: dataware house)와 같은 대용량 데이터 저장소에 쌓이고 있다. 이러한 대규모 데이터 저장소로부터 최적의 의사결정에 필요한 지식을 찾는 과정이 데이터 마이닝이다.

데이터 마이닝의 정의를 간단히 표현하면 대용량 DB로부터 의사 결정에 필요한 지식을 발견(discovery)하는 일련의 과정이다[2]. 좀 더 구체적으로 데이터 마이닝을 정의한다면 크게 다음의 두 가지가 합쳐진 복합적 개념이다.

첫째는 지식의 발견(knowledge discovery)이다. 즉, 데이터를 정보로 바꾸는 숨겨진 패턴(hidden patterns)의 발견이다.

둘째는 지식의 사용(knowledge deployment)이다. 즉, 비즈니스에 도움이 될 수 있게 데이터 마이닝 결과를 지식으로 사용하여 효과적인 의사결정을 수행하는 것이다. 보통은 앞의 지식의 발견만을 데이터 마이닝으로 간주하기도 하는데, 이는 협의의 데이터 마이닝을 정의한 것이다. 광의의 데이터 마이닝은 지식의 사용까지를 포함하는 두 개념의 통합된 정의이다[4].

따라서 데이터 마이닝의 정의는 데이터 마이닝을 연구 및 사용하는 분야에 따라 약간의 차이가 있다. 즉 데이터 마이닝은 이를 다루는 학문 분야에 따라 또는 이러한 솔루션을 개발하는 회사에 따라 약간씩 차이를 보인다. 본 논문에서는 이러한 데이터 마이닝에서 중요한 요소인 마이닝 소프트웨어에 대한 효과적인 사용 전략을 알아본다.

2. 데이터 마이닝 소프트웨어

현재 전세계적으로 사용되는 데이터 마이닝 소프트웨어는 매우 많이 있다. 대표적인 데이터 마이닝 소프트웨어는 SAS 사의 Enterprise-Miner가 있다. 현재 우리나라의 데이터 마이닝 프로젝트의 상당수가 이 소프트웨어를 사용하고 있다. SPSS사의 Clementine도 다양한 데이터 마이닝 분석 기능들을 제공해 준다. 데이터웨어 하우스 전문 회사인 NCR도 테라마이너(tera-miner)라는 마이닝 분석 툴을 제공하고 있다. 이 툴은 특히 자사의 Teradata라는 데이터웨어하우스와 잘 연동된다. 현재 우리나라를 포함하여 전세계적으로 구축되어진 데이터웨어하우스의 상당 부분이 NCR의 Teradata이다.

3. 효과적인 마이닝 전략을 위한 공개 소스시스템

본 논문에서는 현재 국내·외에서 많이 사용되는 5가지의 마이닝 소프트웨어를 비교한다. SAS사의 E-Miner, SPSS사의 Clementine, NCR사의 테라마이너, R, 그리고 마이크로소프트사의 엑셀이다[6],[7],[8],[9]. 물론 엑셀은 스프래드시트 프로그램이지만 간단한 데이터 마이닝 작업에서는 손쉽게 지식을 추출할 수 있는 훌륭한 마이닝 소프트웨어의 역할을 담당할 수 있다. 특히 데이터 마이닝 프로젝트에서 엑셀은 주요 데이터 분석뿐만 아니라 마이닝 결과를 요약하거나 시각화 하는데 중요한 역할을 담당할 때가 많다. 다음 그림은 여러 패키지 제공회사들에 대한 비교이다. 가로축은 알고리즘이 얼마나 강력한가 하는 것을 나타내고 있다. 오른쪽에 위치할수록 알고리즘 성능이 우수하다고 볼 수 있다. 또한 세로축은 툴을 사용에 대한 편리성이다. 위로 올라갈수록 패키지를 사용하기가 어렵다.

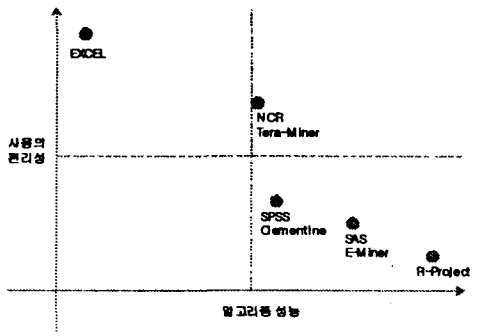

그림 1. 주요 마이닝 소프트웨어 성능비교

위의 그림으로부터 SAS사의 제품은 알고리즘의 성능은 우수하나 사용하기가 어렵다는 것을 알 수 있다. NCR의 제품은 알고리즘의 성능은 보통이나 사용하기는 매우 쉽다는 것을 알 수 있다. 데이터 마이닝 프로젝트에서 사용되는 툴의 선택은 여러 가지 사정들을 모두 고려하여 결정할 수 있다. 위의 제품들 중에는 아주 고가에서부터 저가까지 다양하게 있다. 데이터 마이닝 패키지 프로그램을 구입하여 프로젝트를 수행할 수도 있지만 경우에 따라서는 패키지에서 제공되지 못하는 마이닝 기법도 있게 된다. 이 때에는 수행중인 마이닝 업무를 위한 프로그램을 스스로 개발해서 사용해야 한다. 때문에 마이너는 C와 같은 알고리즘 코딩 언어와 VB(Visual basic)와 같은 포장용 툴에 대한 지식과 사용법도 알고 있으면 좋다. 물론 이러한 작업을 전산 프로그래머에게 부탁할 수도 있지만 데이터 마이닝에 대한 지식이 없는 상황에서 프로그램을 개발한다는 것은 그 만큼 개발 결과에 대한 위험이 따르게 된다. 보통 데이터 마이닝 패키지는 고가이기 때문에 학교에서 마이닝 강의를 위해 구입하기가 어려울 경우에는 R언어 기반의 R의 사용을 추천한다. 이 패키지는 무료이지만 대부분의 데이터 마이닝 알고리즘을 지원하고 있다.

본 논문에서는 효과적인 데이터 마이닝 소프트웨어 사용 전략으로 엑셀과 R의 사용을 제안한다. 기존의 SAS나 SPSS에서 제공하는 데이터 마이닝 소프트웨어는 이들 소프트웨어가 원래 통계학 전공자들을 위한 통계 패키지로 출발했기 때문에 데이터 마이닝 뿐만 아니라 통계학에 대한 기본 지식이 있어야만 효과적인 사용이 가능하다. 또한 이들 소프트웨어는 임대 형식으로 매년 라이센스를 갱신해야 하며 갱신비용이 매우 고가이다. 따라서 중소 데이터 마이닝 프로젝트의 수행이나 통계학을 전공하지 않는 컴퓨터학, 경영학 등의 전공실험실에서 사용하기에는 적절치 않다. NCR의 테라마이너는 NCR의 데이터웨어하우스인 테라데이터가 있어야만 수행되는 문제점이 있다. 테라마이너 자체는 고가가 아니지만 테라데이터가 워낙 고가이므로 이 또한 경제적인 데이터 마이닝 프로젝트 수행에는 적합하지 않다.

반면에 마이크로소프트사의 엑셀은 매우 저렴하다. 하지만 엑셀은 원래 데이터 마이닝 소프트웨어가 아니기 때문에 엑셀만 이용하여 데이터 마이닝 프로젝트를 수행하기에는 어려움이 크다. 하지만 엑셀의 가장 큰 장점은 누구나 간편하게 사용할 수 있고 대부분의 학교, 공공기관, 회사

등에 설치되어 있다는 것이다. 또한 엑셀에서 제공하는 데이터분석 함수와 챠트 마법사를 이용하면 데이터 탐색을 위한 기술통계나 데이터 시각화를 위한 챠트를 매우 손쉽게 제공한다는 것이다.

본격적인 데이터 분석은 R-project를 이용할 수 있다. R은 Bell 연구소 개발된 객체지향형 프로그램 언어인 S 언어가 그 시작이다. 1976년에 개발이 시작된 S 언어는 자료의 조직화, 시각화, 분석을 위하여 개발되었다. 특히, 대화형 시스템의 기반기술의 채택하여 자료의 입·출력 및 변환이 자유로운 데이터 분석용 컴퓨터 언어이다.

S 언어의 설계자인 John Chambers은 S 언어의 개발로 1998년 ACM (Association for Computing Machinery)의 Software System Award 수상하기도 하였다. 그 후 S 언어는 2가지 시스템으로 배포되었다. 그 중 하나인 S-PLUS는 S 언어의 상업용 버전이고 GPL(General Public License)에 따라 CRAN (The Comprehensive R Archive Network)의 홈페이지를 통해 무료로 배포되는 R은 1995년 Auckland 대학(뉴질랜드) Ross Ihaka 교수와 Robert Gentleman 교수에 의해 만들어 졌다.

R이라는 용어는 S 언어에 기반하여 처음으로 개발한 두 교수의 이니셜을 따서 명명하였다. 2002년 R 재단이 발족하여 전 세계의 통계학자, 프로그래머, 그리고 S 언어의 개발자인 John Chambers가 개발에 참여하여 계속적인 개발이 이루어 지고 있다. 이 재단의 R-core 팀은 R을 GNU(Not Unix) S 라고 정의한다. GNU의 GPL을 따르는 공개 소스시스템인 R은 무료로 설치할 수 있고, 소스도 공개되어 있다.

특히, S-PLUS에 비해 자유로운 개발환경과 수많은 라이브러리 제공한다. UNIX, Window, 그리고 Macintosh 등 대부분의 운영체제를 지원하고 있으며 JAVA, Python, C/C++, FORTRAN 등의 컴퓨터 언어와 인터페이스가 가능하며 DBMS의 데이터 접근이 가능한 강력한 마이닝 소프트웨어이다. 이러한 장점 때문에 통계분석 이외에 여러 분야에서 사용되고 있으며 특히, 생물학의 DNA 마이크로어레이(microarray)분석 분야에서는 표준 시스템으로 R이 사용되고 있다 [7].

따라서 본 논문에서 제안하는 효과적인 데이터 마이닝 전략은 다음의 표와 같이 엑셀과 R을 이용한 데이터 분석이다. R은 현재 존재하는 대부분의 기계학습과 통계학 기반의 데이터 분석기법을 제공한다[1],[2],[3],[5].

표 1. 엑셀과 R을 이용한 마이닝.

마이닝 s/w	분석	분석 모듈	
Excel	기술통계	도구-데이터분석(분석도구)	
	시각화	챠트	
R	Regression & Classification	lm(), glm()	선형, 비선형모형
		e1071	SVM, SVR, 베이지안 분류기
		class	K-nearest neighbor SOM, LVQ
		nnet	신경망 모형
	Clustering	svcR	Support Vector Clustering
		e1071	Bagging 군집화
		mva	계층적방법, K-평균
		cluster	계층적 군집화, 퍼지 군집화, 군집화 트리
		SOM	자기조직화 지도
		trimcluster	K-means clustering
	지놈데이터 분석	Bio - conductor	Microarray 자료분석
		limma	Microarray 선형모형

따라서 엑셀과 R를 적절하게 이용한다면 SAS나 SPSS의 상업용 데이터 마이닝 소프트웨어에 의존하지 않고도 얼마든지 효과적인 마이닝 프로젝트를 수행할 수 있게 된다.

4. 결 론

본 논문은 기존에 사용되고 있는 고가의 데이터 마이닝 소프트웨어를 사용하지 않고 엑셀과 공개 소스 소프트웨인 R을 이용하여 효과적으로 마이닝 프로젝트를 수행할 수 있는 전략을 제안하였다. 특히 R은 전세계의 많은 통계학 및 데이터 마이닝 관련 학자, 기술자, 사용자 등의 다양하고 효율적인 개발이 이루어지고 있기 때문에 향후에도 그 사용 범위가 지속적으로 확대되리라 기대된다.

Acknowledgement

This research (paper) was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

참 고 문 헌

[1] A. Ben-Hur, D. Horn, H.T. Siegelmann, and V. Vapnik. A support vector clustering method. in International Conference on Pattern Recognition, 2000.

[2] J. Han, M. Kamber, Data Mining Concept and Techniques, Morgan Kaufmann, 2001.

[3] A.K. Jain and R.C. Dubes. Algorithms for clustering data. Prentice Hall, Englewood Cliffs, NJ, 1988.

[4] NCR, TeraMiner: Data Mining Technology for the Teradata Active Warehouse, Technical White Paper, 2000.

[5] V. N. Vapnik, Statistical Learning Theory, John Wiley & Sons, Inc., 1998.

[6] NCR TeraMiner, www.ncr.com

[7] R Project for Statistical Computing, www.r-project.org

[8] SAS E-miner, www.sas.com

[9] SPSS Clementine, www.spss.com